Fouille de données stochastique pour la compréhension des dynamiques temporelles et spatiales des territoires agricoles. Contribution à une agronomie numérique des territoires

El-Ghali Lazrak

UNIVERSITÉ DE LORRAINE

INRA

T H È S E

pour obtenir le grade de

Docteur

de

l'Université de Lorraine

Spécialité : Sciences agronomiques

présentée et soutenue par

El Ghali LAZRAK

le

19/09/2012

Fouille de données stochastique pour la compréhension des dynamiques temporelles et spatiales des territoires agricoles

Contribution à une agronomie numérique des territoires

Thèse dirigée par **Marc BENOÎT** et **Jean-François MARI**

Unité de recherche : INRA, UR 055, ASTER-Mirecourt, F-88500, Mirecourt

École doctorale : Sciences et ingénierie des Ressources, Procédés, Produits, Environnement (RP2E)

Composition du jury

Peter VERBURG, Professeur, **VU University Amsterdam**	**Rapporteur**
Florence FORBES, Directrice de recherche, **INRIA**	**Rapporteur**
Frédérick GARCIA, Directeur de recherche, **INRA**	**Examinateur**
Jean-Marc MEYNARD, Directeur de recherche, **INRA**	**Examinateur**
Amedeo NAPOLI, Directeur de recherche, **CNRS**	**Examinateur**
Marc BENOÎT, Directeur de recherche, **INRA**	**Directeur de thèse**
Jean-François MARI, Professeur, **Université de Lorraine**	**Directeur de thèse**

Remerciements

Cette thèse est le résultat d'un travail de si longue haleine qui ne m'aurait pas été possible de le mener à bout sans le soutien d'hommes et de femmes qui ont été à mon sens un modèle de patience et de compétence. Je ne saurais trouver les mots justes pour les remercier de leur dévouement, de leur disponibilité et de leur tolérance.

En disant ces mots je pense très fort à :

- Marc et Jean-François mes encadrants durant ces années de recherche qui, malgré leurs nombreuses charges m'ont supporté avec la même égale compréhension surtout dans mes moments difficiles. L'interaction avec vous était très enrichissante.
- Catherine Mignolet, la directrice de l'unité pour ses remarques scientifiques avisées et pour m'avoir facilité au maximum les démarches administratives nécessaires au déroulement de mon travail.
- Jean-Marie Trommenschlager et Damien Foissy pour leur disponibilité et leur efficacité à résoudre mes soucis informatiques.

Mes remerciements vont également :

- Aux membres du jury : Amedeo Napoli, Florence Forbes, Frédéric Garcia, Jean Marc Meynard et Peter Verburg pour l'intérêt qu'ils ont montré à l'égard de mon travail. Je remercie tout particulièrement Florence Forbes et Peter Verburg qui ont bien voulu en être les rapporteurs.
- À mon comité de pilotage : Vincent Bretagnolle, Florence Le Ber, Thomas Houet et Philippe Martin grâce à qui j'avais pu voir plus clair après chaque réunion.
- À Noémie Schaller pour sa fructueuse collaboration.
- À Davide Rizzo pour l'intérêt que tu m'as constamment manifesté pour lire ma thèse, ce qui me motivait à avancer. Tes remarques sur le premier chapitre m'étaient précieuses.

Je remercie Mylène Voiry et Florian Beller d'avoir accepté de réaliser leurs stages avec moi et je leur dis que leurs travaux m'ont permis d'avancer.

Je souhaite un bon bouclage de leurs thèses à Ying, Laura et Kamal et les remercie pour le coup de main apporté à la préparation du pot de ma soutenance.

Je veux aussi remercier mes parents, ma sœur Zineb et Khaled son mari pour leur soutien indéfectible tout le long de mes années de thèse et de mes études antérieures. Un merci tout particulier à papa pour sa lecture attentionnée de la dernière version de ce document et les corrections de syntaxe. Merci à ma petite famille pour sa compréhension de mes absences pour le travail.

Je suis loin d'avoir cité dans mes remerciements toutes les personnes qui ont contribué de près ou de loin à mon travail et m'excuse auprès de ceux, nombreux, que je n'ai pas désigné nommément.

Table des matières

Liste des abréviations **v**

Chapitre 1 Introduction et problématique **1**

1.1 Concepts clés de l'OTAA 2
1.2 L'OTAA détermine les enjeux environnementaux en milieu agricole 4
1.3 Spécificités de l'OTAA 5
1.4 Deux communautés de chercheurs pour un objet commun : le territoire agricole 6
1.5 Notre représentation conceptuelle de l'OTAA 7
1.6 Contexte de la thèse et question de recherche 8
1.7 Références 10

Chapitre 2 Corpus de données spatio-temporelles d'OCS et outils de fouille utilisés **17**

2.1 Territoires d'étude et corpus de données 17
2.2 Fouille de données spatio-temporelles d'OCS par segmentation à l'aide de HMM2 22
2.3 Présentation et mise en œuvre d'ARPENTAGE 34
2.4 Conclusion 44
2.5 Références 46

Chapitre 3 Notre méthode de modélisation de l'OTAA avec ARPENTAGE **49**

3.1 Statistiques sur le corpus de données 50
3.2 Segmentation temporo-spatiale 52
3.3 Fouille des voisinages des OCS et des successions d'OCS 62
3.4 Interactions avec les experts 68
3.5 Conclusion 71
3.6 Références 72

Chapitre 4 Segmentation temporo-spatiale **75**

4.1 Landscape regularity modelling for environmental challenges in agriculture 76
4.2 Segmentation temporo-spatiale des successions de cultures d'un territoire agricole à l'aide de HMM2 91
4.3 Segmentation temporo-spatiale du bassin versant du Yar fondée sur des occupations du sol télédétectées 105

Chapitre 5 Analyse des relations de voisinage des successions d'OCS 113
5.1 Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales . . 115
5.2 Fouille de paysages agricoles : analyse des voisinages des successions d'occupation du sol 122

Chapitre 6 Articulation des régularités stochastiques avec les règles de décision 129
6.1 Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol 130
6.2 Combining farmers' decision rules and landscape stochastic regularities for landscape modelling 142

Chapitre 7 Discussion Générale 157
7.1 Retour sur les hypothèses 157
7.2 Apports, limites et perspectives d'amélioration des méthodes proposées 162
7.3 Pour conclure : contribution à l'agronomie numérique des territoires 166
7.4 Références 167

Références générales 177

Annexe A Using Markov Models to Mine Temporal and Spatial Data 179

Liste des abréviations

ADN	Acide désoxyribonucléique
ANR	Agence nationale de la recherche
CEBC	Centre d'Études Biologiques de Chizé
CEM	Classification EM
Costel	Climat et occupation du sol par télédétection
EA	Exploitation agricole
EM	Expectation-maximisation algorithm (algorithme espérance-maximisation)
GLP	Global Land Project
HHMM2	Second Order HHMM (HHMM d'ordre 2)
HHMM	Hierarchical HMM (HMM hiérarchique)
HMM1	First Order HMM (HMM d'ordre 1)
HMM2	Second Order HMM (HMM d'ordre 2)
HMM	Hidden Markov Model (modèle de Markov caché)
IGBP	International Geosphere-Biosphere Programme
IHDP	International Human Dimensions Programme
LCS	Land Change Science
LUCC	Land Use and Land-Cover Changes
Loria	Laboratoire lorrain de recherche en informatique et ses applications
MAP	Maximum *a posteriori* probability
NP	Non Polynomial
OCS	Occupation du sol
OTAA	Organisation territoriale de l'activité agricole
PAC	Politique agricole commune
PMI	Pointwise Mutual Information (Information mutuelle spécifique)
PRA	Petite région agricole
SIG	Système d'information géographique
UAP	Unité agro-physionomique
UA	Unité agronomique

Chapitre 1

Introduction et problématique

Le changement global de l'environnement s'impose aujourd'hui comme une préoccupation majeure qui menace les ressources naturelles et compromet les générations futures. Ce changement résulte d'une grande variété d'activités humaines parmi lesquelles l'agriculture est celle qui utilise et transforme la plus grande partie de la surface terrestre (VITOUSEK et al., 1997 ; FAOSTAT, 2009). La science du changement des territoires (LCS pour Land Change Science) a récemment émergé comme une branche multidisciplinaire faisant partie intégrante de la recherche sur le changement global de l'environnement. La LCS s'intéresse au système « Anthropisation - Environnement » en particulier aux interactions qui conduisent aux dynamiques de l'occupation et de l'utilisation du sol. Elle cherche à comprendre ces dynamiques, leurs forces motrices et leurs conséquences (RINDFUSS et al., 2004 ; TURNER et al., 2007). La LCS s'est concrétisée avec les projets de recherche LUCC (pour Land Use and Land-Cover Changes : 1995-2005) et GLP (pour Global Land Project : 2005-en cours) issus du couplage des efforts des programmes IGBP (International Geosphere-Biosphere Programme) et IHDP (International Human Dimensions Programme on Global Environmental Change).

Les travaux liés au projet LUCC étaient principalement orientés vers l'élaboration d'une nouvelle génération de modèles capables de simuler les principales forces motrices socio-économiques et biophysiques du changement de l'utilisation et de l'occupation des sols (VERBURG et al., 2006). Le projet LUCC s'est focalisé sur 3 axes de recherche (TURNER et al., 1995) :

1. l'analyse comparée d'études de cas de la dynamique d'utilisation des sols ;
2. l'observation empirique et l'élaboration de modèles diagnostiques de la dynamique de couverture des sols ;
3. le développement de modèles intégrés régionaux et globaux.

Depuis la fin du projet LUCC en 2005, le projet GLP a pris la relève en s'appuyant sur les acquis de son prédécesseur et en s'ouvrant davantage sur les sciences humaines et sociales (*GLP* 2005). Le projet GLP focalise sur 3 objectifs qui consistent à :

1. identifier les agents, les structures et la nature du changement dans les systèmes « Anthropisation - Environnement » sur les terres, et quantifier leurs effets sur ces systèmes ;
2. évaluer la façon dont la fourniture de services écosystémiques est affectée par ces changements ;
3. identifier le caractère et la dynamique vulnérables et durables aux perturbations interagissant avec ces systèmes, y compris le changement climatique.

Parmi les travaux réalisés sur les dynamiques de l'occupation et de l'utilisation du sol, ceux centrés sur les territoires agricoles sont relativement peu nombreux et sont souvent menés par des équipes multidisciplinaires où l'agronomie fait défaut. Un point commun à la plupart de ces travaux consiste à considérer les cultures annuelles comme un type unique de couverture du sol. Du point de vue des agronomes, cette vision gagnerait à être approfondie en vue de pouvoir considérer l'utilisation des territoires agricoles et leurs dynamiques plutôt que la simple considération des changements de la couverture agricole du sol.

Cette thèse s'insère dans ce courant de recherche et s'intéresse en particulier aux dynamiques passées et actuelles de l'utilisation des territoires [1] agricoles de dimensions régionales. Ces dynamiques relèvent soit du fonctionnement régulier de l'activité agricole soit de changements dans son fonctionnement. Pour modéliser ces deux cas de figures, il s'agira dans ce travail de modéliser l'Organisation Territoriale de l'Activité Agricole (OTAA).

Ces aspects sont repris avec plus de détail dans la suite de ce chapitre. Mais avant de les développer, marquons un temps d'arrêt sur des concepts clés que nous mobilisons pour nos recherches sur l'OTAA.

1.1 Concepts clés de l'OTAA

1.1.1 L'assolement, la rotation et la succession

L'assolement « est la répartition des cultures entre les parcelles d'un même finage [2], ou entre les soles [3] si le finage connaît la contrainte des soles. L'assolement se justifie parce qu'on cherche à obtenir le meilleur rendement en épuisant le moins possible la terre ... » (LACHIVER, 1997). Selon la même source, un assolement que peut encore pratiquer l'agriculture d'aujourd'hui était déjà pratiqué dans l'antiquité : « ...l'assolement biennal, bien attesté dans le monde romain, peut se pratiquer avec ou sans jachère, une culture de fève, par exemple, pouvant suivre un blé d'hiver ... ». Le concept d'assolement, où s'entremêlent raisonnements temporel et spatial opéré par l'agriculteur, n'a pas nécessité de profondes évolutions pour s'adapter à l'agriculture d'aujourd'hui, mais il s'est complexifié afin de répondre à de nouvelles contraintes : comme la taille de l'exploitation agricole, sa structure, la nature du sol des parcelles, etc. Selon le dictionnaire d'agriculture (1999), l'assolement est « la répartition pendant une année donnée des terres labourables d'une exploitation

1. Par ***territoire***, nous entendons : un espace de dimension régionale géré par un groupe d'acteurs.
2. ***Finage*** : ensemble des terres exploitées par une même communauté rurale.
3. ***Sole*** : unité de l'assolement formée de parcelles groupées en quartiers et consacrées à la même culture ou à la jachère.

agricole entre les différentes cultures ». L'assolement est aussi « la terre sur laquelle on a pratiqué l'assolement » (REY, 2005).

La rotation est définie dans le *Dictionnaire du monde rural : les mots du passé* de LACHIVER (1997) par « l'alternance des cultures sur un champ selon le rythme annuel, mais parfois saisonnier quand il s'agit de cultures légumières ... ». Selon le dictionnaire d'agriculture (1999) la rotation est l'ordre de succession sur la même parcelle, des plantes appartenant à des espèces ou à des variétés différentes et éventuellement de jachères, cette succession se répétant régulièrement dans le temps.

La succession de cultures est l'enchainement temporel, non toujours régulier, des couverts végétaux dans une parcelle culturale (SEBILLOTTE, 1990).

Le dictionnaire d'agriculture (1999) note qu'il est souvent confondu entre assolement et rotation dans certaines expressions qui qualifient cette dernière : assolement biennal, triennal au lieu de rotation biennale, triennale. Nous pensons que cette confusion terminologique est due au fait que la façon dont chaque agriculteur organise son territoire d'exploitation est un processus à la fois temporel et spatial.

1.1.2 Le système de culture

Le système de culture se définit au niveau de chaque parcelle ou ensemble de parcelles traitées de façon homogène, par

1. la succession de cultures, et pour chaque culture, par
2. les itinéraires techniques qui sont des combinaisons logiques et ordonnées d'opérations culturales permettant de contrôler les états du milieu et des populations végétales cultivées pour en tirer une production donnée (SEBILLOTTE, 1990 ; BENOÎT et PAPY, 1998).

Un territoire agricole peut être vu comme un ensemble de portions d'espace affectées à des systèmes de culture (BENOÎT et PAPY, 1998). Dans notre thèse, nous adoptons ce point de vue et considérons le système de culture comme un concept clé pour modéliser l'OTAA en vue de contribuer à l'étude des relations entre l'activité agricole et les questions environnementales. Nous insisterons sur l'étude des successions de cultures et ne traiterons pas des itinéraires techniques.

1.1.3 La succession de cultures comme « brique de base » de l'OTAA

La succession de cultures est une notion plus large que celle de la rotation, qu'elle inclut. Elle traduit l'adaptation des agriculteurs aux contraintes et opportunités qu'ils rencontrent entre leurs projets de production et les caractéristiques de leurs territoires d'exploitation. En temps de stabilité agricole, les successions de cultures correspondent souvent à des rotations.

Lorsqu'il s'agit d'étudier l'OTAA à des échelles régionales englobant d'une centaine à plusieurs milliers de territoires d'exploitations agricoles, il devient difficile d'obtenir de l'information sur les itinéraires techniques (deuxième composante du concept de système de culture) car ceux-ci nécessitent de

réaliser des enquêtes au niveau de l'exploitation agricole avec une consultation des archives ou de la mémoire de l'agriculteur. Par contre, l'information sur les successions de cultures (première composante du concept de système de culture) est plus facilement accessible que ce soit à travers des relevés réguliers de l'occupation et de la configuration des parcelles du territoire d'étude, ou par le traitement d'images satellites. Dans le présent travail, l'activité agricole sera décrite par les successions de cultures supposées capables de décrire l'OTAA de manière utile pour les spécialistes des questions environnementales étudiées. Ce rôle clé des successions de cultures a été mis en avant pour différentes questions environnementales : comme la qualité de l'eau (MIGNOLET et al., 2004 ; HERRERA et LIEDGENS, 2009), la biodiversité (BAUDRY et al., 2003 ; BUREL et BAUDRY, 2005) ou l'érosion (MARTIN et al., 1998).

1.1.4 De la succession de cultures à la succession d'occupation du sol

Dans un territoire agricole de dimension régionale se mêlent activité agricole, zones urbaines, réseau routier, forêts, réseau hydrographique, etc. Ainsi, les terres agricoles peuvent êtres bâties ou boisées. Afin de tenir compte de cette diversité de couverts et de leurs dynamiques, la succession d'occupation du sol (OCS) sera désormais utilisée pour décrire la dynamique ou la stabilité des couvertures du sol dans le territoire agricole.

1.2 L'OTAA détermine les enjeux environnementaux en milieu agricole

L'utilisation agricole des territoires constitue une part importante de l'utilisation du sol tant au niveau global qu'aux niveaux européen et français (tableau 1.1).

TABLEAU 1.1 – Parts respectives des terres arables et des terres agricoles aux niveaux mondial, européen et français (FAOSTAT, 2009).

	Terres arables [a] (%)	Terres agricoles [b] (%)
Monde	10,3	36,3
Union Européenne	25,1	43,5
France	33,4	53,3

a. Les ***terres arables*** comprennent les terrains en jachère, les cultures maraîchères et céréalières et les prairies artificielles.
b. ***Terres agricoles*** incluent, en plus des terres arables, les terres sous cultures permanentes (arboriculture et viticulture) et les prairies naturelles.

Au niveau européen, l'activité agricole s'est intensifiée, depuis 1962, sous l'influence de la Politique Agricole Commune (PAC) et l'impact négatif de cette intensification sur l'environnement n'est plus à démontrer (VITOUSEK et al., 1997 ; ORMEROD et WATKINSON, 2000 ; DONALD et al., 2001 ; COPPEDGE et al., 2001 ; LUNT et SPOONER, 2005 ; TSCHARNTKE et al., 2005 ; INCHAUSTI et BRETAGNOLLE, 2005 ; AVIRON et al., 2007 ; RAINI, 2009).

Pour les agronomes l'enjeu est de se placer dans un niveau d'organisation où ils cherchent à comprendre les liens entre activités et territoires agricoles. Avec les premières approches globalisantes dans

les années 60, l'agronomie s'intéresse au niveau de référence qu'est la parcelle, aux caractéristiques et fonctionnement du sol (HÉNIN, 1999). Ce type d'approche se développant et les niveaux d'analyse changeant, il est ensuite fait référence au territoire à des échelles régionales. Il peut s'agir de diagnostics régionaux des potentialités agricoles (HÉNIN et DEFFONTAINES, 1970) ou de démarches monographiques autour de terroirs (BONNEMAIRE, 1995). Parallèlement, dans les années 70, se développent des travaux autour d'un nouveau niveau d'analyse : l'exploitation agricole vue comme un système (OSTY, 1978). Dans ce cadre se développent des recherches autour de la prise de décision (SEBILLOTTE et SOLER, 1988 ; SEBILLOTTE et SOLER, 1990) puis autour des relations qui lient l'espace aux activités agricoles (LANDAIS et DEFFONTAINES, 1990 ; DEFFONTAINES et al., 1995). Aujourd'hui, des agronomes faisant du territoire agricole leur niveau de référence (SEBILLOTTE, 2005), contribuent à construire des systèmes de production agricole durables à l'échelle des territoires. Ces agronomes se positionnent dans une nouvelle perspective agronomique : l'***agronomie des territoires*** (BENOÎT et al., 2007 ; BENOÎT et al., Soumis ; MÉROT et al., Soumis). C'est dans ce courant de recherche que s'inscrit cette thèse.

Pour un problème environnemental identifié, la description du système est une étape clé dans la séquence d'analyse et du traitement des problèmes environnementaux liés aux changements de la couverture et de l'utilisation du sol (STOMPH et al., 1994 ; VERBURG et VELDKAMP, 2001). Dans un territoire à dominante agricole, cette étape consiste à décrire les dynamiques temporelles et spatiales des systèmes de cultures. Dans la littérature, la description des dynamiques temporelles et spatiales des systèmes de cultures souffre souvent d'un manque de considération des spécificités temporelles et spatiales de l'activité agricole. Dans notre travail de thèse, nous nous intéressons à cette étape à travers la modélisation de l'OTAA.

1.3 Spécificités de l'OTAA

Une partie souvent non négligeable de l'espace agricole est occupée par des cultures annuelles qui, d'année en année, peuvent apparaître comme des changements de l'OTAA alors qu'en réalité ce sont de simples réarrangements spatiaux opérés par les agriculteurs lorsqu'ils assolent [4] les terres de leurs exploitations. D'autre part, un changement de système de culture peut se faire sans modification des surfaces en cultures annuelles, rendant impossible le diagnostic de ce changement sans la prise en compte des successions dans lesquelles s'insèrent ces cultures annuelles. Ces dépendances entre l'espace et le temps sont d'autant plus complexes lorsqu'il s'agit de territoires de dimensions régionales où la mosaïque agricole est construite par de nombreux acteurs. Ces acteurs sont soumis à des opportunités et des contraintes de nature différente agissant à diverses échelles.

L'ensemble de ces contraintes/opportunités nous fait émettre l'hypothèse que l'activité des différents agriculteurs, sans être concertée, conduit à l'organisation de l'espace agricole d'une manière convergente et intelligible qu'il convient d'exploiter dans notre modélisation de l'OTAA pour simplifier sa complexité.

4. ***Assoler*** consiste à diviser les terres d'une exploitation en soles sur chacune desquelles on fait alterner les cultures dans un ordre déterminé.

Soulignons que dans cette thèse nous nous sommes intéressés au diagnostic des dynamiques passées et actuelles de l'utilisation de territoires agricoles de dimensions régionales. Les travaux avec une optique prospective comme celui de HOUET (2006) ou de simulation de paysages agricoles comme ceux de CASTELLAZZI et al. (2008) et DOGLIOTTI et al. (2003) n'entrent pas dans le cadre de cette thèse.

1.4 Deux communautés de chercheurs pour un objet commun : le territoire agricole

Agricultures et territoires sont des termes de plus en plus abordés par les agronomes et les géographes. Les agronomes s'intéressent à des territoires plus larges que la parcelle, et les géographes, seuls ou en collaboration avec d'autres disciplines, s'investissent sur la question de l'organisation de l'activité agricole dans des territoires porteurs d'enjeux environnementaux. Pourtant, l'objet d'étude partagé ne signifie pas que les échanges s'intensifient entre ces deux communautés (BÜHLER et al., 2010). Les géographes s'intéressent au diagnostic, sur des longs ou moyens termes, des dynamiques d'occupation et d'utilisation du sol dans les territoires agricoles (SIMPSON et al., 1994 ; POUDEVIGNE et ALARD, 1997).

Un exemple typique de cette approche est celui de SIMPSON et al. (1994) qui ont mené une étude comparative, sur l'évolution en 48 ans, de deux territoires contigus initialement à dominante agricole couvrant un total 242 km^2. En s'appuyant principalement sur 4 photos aériennes prises à des pas de temps de l'ordre de la décennie, ces auteurs distinguent 4 catégories d'ocs : l'agriculture, la forêt, l'urbain et l'industrie. La comparaison de l'évolution des deux territoires est quantitative en termes de progression et de régression de ces 4 ocs. Dans un même cadre méthodologique, POUDEVIGNE et ALARD (1997) ont travaillé sur le bassin versant de Brionne (Normandie) qui s'étend sur une trentaine de km^2. Ces auteurs se sont intéressés à décrire l'organisation du bassin versant de Brionne et de son évolution sur une période de 25 ans. En s'appuyant sur deux photos aériennes correspondant au début (1964) et à la fin (1989) de la période d'étude, ils distinguent 5 ocs : l'urbain, la prairie, la forêt, les vergers, et les cultures. Les évolutions de ces ocs sont estimées entre le début et la fin de la période d'étude. Les évolutions d'ocs constatées ont ensuite été reliées à une typologie d'exploitations agricoles réalisée avec une analyse multivariée.

Ce type d'approche consiste à considérer les cultures annuelles (blé, orge, tournesol, colza, etc.) comme une catégorie unique de couverture du sol et de chercher à identifier les changements quantitatifs à travers une succession d'images affectant cette catégorie agricole sur du long à moyen terme avec des pas de temps pluriannuels voire multi-décennaux. Avec ce type d'approche, les spécificités agronomiques comme l'effet précédent-suivant, le temps de retour d'une culture sur elle-même ne sont pas pris en compte. En outre, les changements diagnostiqués peuvent ne pas en être, par exemple la transition « prairies temporaires-blé » peut être une succession logique en terme agronomique.

En accord avec de nombreux auteurs (PAPY et TORRE, 2002 ; CREWS-MEYER, 2004 ; MARTIN et al., 2006 ; BENOÎT, 2006 ; MIGNOLET et al., 2007 ; SCHIETTECATTE et al., 2008 ; THENAIL et al., 2009) nous considérons que les études des trajectoires sur le long terme gagneraient à être complétées

par une description plus détaillée de l'OTAA. La représentation de l'OTAA doit permettre de repérer à la fois les dynamiques cycliques liées aux systèmes de cultures pratiqués dans le territoire, et les éventuelles périodes de rupture qui reflètent des changements dans les pratiques agricoles. Ces deux composantes de l'OTAA sont utiles pour la recherche de liens entre l'activité agricole et les dynamiques liées à l'enjeu environnemental en question. A l'échelle de la décennie, il est généralement possible d'apprécier ces dynamiques et de différencier entre les dynamiques inhérentes aux rotations de cultures et les dynamiques dues à des changements de systèmes de cultures.

Les travaux s'inscrivant dans l'agronomie des territoires comme (MIGNOLET et al., 2007 ; MOTTET et al., 2006), et les travaux des géoagronomes (THINON et DEFFONTAINES, 1999 ; DEFFONTAINES et THINON, 2001 ; THINON, 2003 ; JACOPIN, 2011) sont des exemples de collaboration fructueuse entre agronomie et géographie qui ont permis de modéliser l'OTAA de façon élaborée mais ces travaux restent peu nombreux.

DEFFONTAINES et THINON (2001) ont représenté l'OTAA à des échelles régionales en définissant, par une méthode visuelle, des portions de territoires de relative égale apparence : les Unités Agro-Physionomiques (UAP). La cartographie des UAP est fondée sur l'hypothèse stipulant qu'il existe dans l'espace rural des portions de territoires, dénommées Unités Agronomiques (UA), qui sont de taille nettement supérieure à celle des parcelles agricoles, et qui présentent une relative égale organisation des usages agricoles. Dans une seconde hypothèse ces auteurs supposent l'existence d'une relation entre l'UA et le concept de système de culture. Selon cette hypothèse, les systèmes de culture sont organisés dans le territoire en UA sous l'influence de champs géographiques diversifiés et simultanément, ces systèmes de cultures contribuent à la délimitation des UAP par les formes agraires qui caractérisent chacun d'eux. La cartographie des UAP réalisée sur divers espaces agricoles est une concrétisation de la notion d'UA et constitue une validation experte des UA. Toutefois, la démarche aboutissant au partitionnement du territoire agricole en UAP reste limitée en raison de :

1. la nécessité d'une double expertise en agronomie et en géographie, et
2. du caractère subjectif dû à l'adoption d'une méthode visuelle.

Ces travaux ne sont pas reproductibles à des échelles régionales car ils s'appuient fortement sur du savoir-faire d'experts. Nous gardons les mêmes hypothèses fondatrices des UAP et nous tentons de mettre au point une méthode informatique de reconnaissance des unités de territoires ayant une homogénéité de dynamiques agronomiques.

1.5 Notre représentation conceptuelle de l'OTAA

La façon dont chaque agriculteur organise son territoire d'exploitation est un processus à la fois temporel et spatial (1.1). Nous faisons l'hypothèse que cette interdépendance entre le temps et l'espace se maintient à une échelle englobante de territoires agricoles de dimensions régionales. Ceci nous permet de formuler 3 hypothèses :

1. les successions de cultures sont relativement courtes et organisées dans le temps. En d'autres termes, la culture d'une année dépend des cultures au même endroit de quelques années précédentes ;
2. l'agriculteur alloue une culture à une parcelle en tenant compte du voisinage de cette parcelle. En d'autres termes, les cultures sont organisées dans l'espace et dépendent de leur voisinage proche ;
3. ces deux hypothèses jointes, suggèrent une 3e hypothèse : le choix d'une succession de cultures en un point de l'espace dépend du choix des successions de cultures voisines.

Ces hypothèses décrivent la complexité de l'OTAA par les dépendances des voisinages temporels et spatiaux, et confèrent à l'OTAA une propriété locale, appelée propriété de Markov, selon laquelle l'évolution du processus à un instant ou (et) un lieu donné(s) est uniquement déterminée par les valeurs du voisinage temporel ou (et) spatial. La première hypothèse assimile la succession de cultures à une chaîne de Markov, la seconde hypothèse assimile l'espace agricole à un champ de Markov, et la troisième hypothèse assimile l'organisation temporelle et spatiale des successions de cultures à un couplage de chaînes de Markov et de champs de Markov qui s'opère dans l'espace et dans le temps. Ceci nous mène à représenter l'OTAA comme une image de successions plutôt que comme une succession d'images.

1.6 Présentation des contextes scientifiques et institutionnels de la thèse et question de recherche

Cette thèse s'inscrit dans la continuité d'une série de travaux de modélisation stochastique de l'OTAA menés depuis 1998 par mon équipe d'accueil à l'INRA ASTER [5] de Mirecourt en étroite collaboration avec le Loria [6]. Ces travaux ont montré que les modèles de Markov cachés [7] permettent d'identifier et de quantifier les successions de cultures pratiquées par les agriculteurs dans des territoires agricoles de dimensions régionales (MARI et al., 1998 ; MARI et al., 1999), et de décrire leur répartition spatiale et leur évolution temporelle (MIGNOLET et al., 2004 ; LE BER et al., 2006 ; MARI et LE BER, 2006). Initialement, ces modèles étaient utilisés sur des données spatio-temporelles d'occupation du sol (OCS) issues des enquêtes *Teruti* [8] du ministère de l'agriculture. La méthode développée dans cette thèse vise à systématiser la modélisation de l'OTAA sur des données moins bien calibrées que les données *Teruti* ainsi que sur des territoires agricoles de différentes tailles. Nous avons travaillé pour cela sur deux territoires d'études de tailles et de problématiques environnementales différentes dotés de corpus de données spatio-temporelles d'OCS pour l'un issu de relevés de terrain et pour l'autre issu de la télédétection.

5. Agro-Sytèmes TErritoires Ressources.
6. Laboratoire lorrain de recherche en informatique et ses applications.
7. ***Les modèles de Markov cachés*** seront présentés plus loin (section 2.2 page 22).
8. ***L'enquête Teruti*** , réalisée chaque année depuis 1982 par les services statistiques du ministère de l'alimentation, de l'agriculture et de la pêche, permet de suivre l'évolution des différentes catégories d'OCS à partir d'un ensemble de points constituant un échantillon représentatif du territoire national. Chaque année, les enquêteurs relèvent l'OCS sur l'ensemble de ces points. Ils leur attribuent un code dans une nomenclature de 81 modalités d'OCS agricoles et non agricoles. L'échantillon *Teruti* a été modifié à plusieurs reprises pour corriger des biais statistiques et pour l'étendre aux départements d'outre-mer.

Dix ans après la naissance de cette collaboration entre agronomes des territoires et informaticiens spécialistes en fouille de données, l'Agence Nationale de la Recherche (ANR) permettait le financement du projet BiodivAgriM [9] qui a mobilisé plusieurs thèses cherchant :

1. à mieux comprendre et évaluer les conséquences sur la biodiversité des **mosaïques paysagères générées par l'agrégation des systèmes de culture des exploitations agricoles**,
2. et à repérer les points critiques et les leviers d'action dans les réorganisations des exploitations agricoles.

Cette thèse — effectuée dans ce cadre — cherchait à contribuer à la compréhension du premier point en essayant de répondre à la question de recherche suivante :

Comment modéliser l'OTAA à l'échelle de territoires agricoles régionaux compatibles avec les échelles où s'expriment les services environnementaux et écologiques ?

Pour y répondre, nous avons mis au point et expérimenté **une méthode informatisée de modélisation de l'OTAA, reproductible à partir de données de sources variées (relevés de terrain, télédétection), permettant de fouiller un corpus de données spatio-temporelles d'OCS en vue d'identifier les dynamiques agricoles qui s'y opèrent.**

Organisation du manuscrit

La suite de ce document est organisée en 6 chapitres :

Le chapitre suivant présente les territoires d'étude et décrit les techniques et les outils informatiques que nous avons mobilisés pour modéliser l'OTAA.

Le chapitre 3 présente une synthèse de notre méthode de modélisation de l'OTAA. Ce chapitre reprend tous les éléments méthodologiques mobilisés dans nos articles. Cette présentation méthodologique est complétée par les 3 chapitres suivants qui reprennent tous nos travaux publiés entre 2008 et 2011.

Le chapitre 4 présente notre procédure de fouille de données permettant la segmentation temporo-spatiale d'un territoire agricole de dimension régionale à travers l'extraction de régularités temporelles puis leur spatialisation sous forme d'une carte où les unités cartographiques sont décrites par les distributions probabilistes de ces régularités. Cette procédure de fouille de données est illustrée à travers trois articles :

9. Le projet BiodivAgriM (2008-2011) est intitulé « conservation de la biodiversité dans les agro-écosystèmes : une modélisation spatialement explicite des paysages ». Ce projet visait à articuler les dynamiques d'organisation interne des exploitations agricoles, les dynamiques d'organisation spatiale des paysages et les dynamiques de fonctionnement écologique des agro-écosystèmes, pour identifier les leviers d'action dans les réorganisations des exploitations agricoles. L'exploitation agricole étant un niveau majeur de décision et d'organisation, mais les services environnementaux et écologiques s'expriment à d'autres échelles plus vastes (paysages, bassin versant, ...). Cette discordance observée entre les niveaux d'organisation des processus de gestion des territoires agricoles et les niveaux d'organisation des processus écologiques liés au maintien de la biodiversité constitue une difficulté majeure traitée par ce projet.

- Le premier article (LAZRAK et al., 2009) présente la procédure de fouille de données appliquée au site d'étude de Chizé. Pour ce cas d'étude, la fouille de données a porté sur un corpus d'ocs construit par des relevés de terrain effectués le long de la période d'étude par des enquêteurs du CNRS à Chizé. La fouille de données a permis l'extraction des successions dominantes et leur spatialisation sous forme d'une carte de patches de successions.
- Le deuxième article (LAZRAK et al., 2009), adressé à un public d'informaticiens, reprend les principales étapes de la procédure de fouille temporo-spatiale sans développer les aspects agronomiques.
- Le troisième article, en préparation, présente un test de généricité de la procédure de fouille initialement développée pour le cas du site d'étude de Chizé. Ce test de généricité a permis d'étendre l'utilisation de cette procédure de fouille pour une série d'assolements annuels construits à partir d'images satellites.

Le chapitre 5 présente notre procédure de fouille de données permettant d'analyser les dépendances entre les successions d'ocs voisines à l'échelle d'un territoire agricole de dimension régionale. Nous avons appliqué cette procédure de fouille au site d'étude de Chizé à travers deux articles :

- Le premier article (LAZRAK et al., 2010) propose de modéliser la variabilité temporelle et spatiale de la mosaïque agricole.
- Le deuxième article (MARI et al., 2010) explore la complémentarité de deux approches d'analyse des relations de voisinages entre successions d'ocs dans un territoire agricole représenté par sa mosaïque de parcelles.

Le chapitre 6 présente deux articles illustrant nos interactions avec des experts agronomes dans un cadre d'analyse articulant deux approches complémentaires : la modélisation des régularités stochastiques sur les dynamiques de voisinage des ocs, et la modélisation des règles de décisions d'agriculteurs identifiées par enquêtes. Ce chapitre comporte deux articles :

- Le premier article (LAZRAK et al., 2011), destiné pour un public d'informaticiens, explique la méthode proposée en développant la méthode de fouille de données sans développer les aspects agronomiques.
- Le deuxième article (SCHALLER et al., 2011) développe la méthode proposée du point de vue de l'agronome des territoires, en mettant l'accent sur la complémentarité entre les deux approches ainsi que sur les limites et perspectives d'amélioration de cette méthode.

Nous clôturons ce manuscrit par une **discussion générale** des apports, des limites et des perspectives d'amélioration de la méthode de modélisation proposée.

1.7 Références

AVIRON, S, P KINDLMANN et F BUREL (2007). « Conservation of butterfly populations in dynamic landscapes: The role of farming practices and landscape mosaic ». Dans : *ecological modelling* 205.1-2, p. 135–145.

BAUDRY, J, F BUREL, S AVIRON, M MARTIN, A OUIN, G PAIN et C THENAIL (2003). « Temporal variability of connectivity in agricultural landscapes: do farming activities help? » Dans : *Landscape ecology* 18.3, p. 303–314.

BENOÎT, M (2006). « Organisation territoriale des activités agricoles ». Dans : *Acteurs et territoires locaux: vers une géoagronomie de l'aménagement*. Editions Quae, p. 87–89.

BENOÎT, M et F PAPY (1998). « La place de l'agronomie dans la problématique environnementale ». Dans : *Dossiers de l'environnement INRA* 17, p. 53–71. URL : http://www.inra.fr/dpenv/benoid17.htm.

BENOÎT, M, C MIGNOLET, S HERRMANN, D RIZZO, C MOONEN, P BARBERI, M GALLI, E BONARI, N SILVERSTRI, C THENAIL, S LARDON, H RAPEY, E MARRACCINI, F LE BER et JM MEYNARD (2007). « Landscape designed by farming systems: a challenge for landscape agronomists in Europe ». Dans : *Farming Systems Design 2007, Methodologies for Integrated Analysis of Farm Production Systems*. Catania, Sicilia, Italy, p. 2.

BENOÎT, M, D RIZZO, E MARRACCINI, AC MOONEN, M GALLI, S LARDON, H RAPEY, C THENAIL et E BONARI (Soumis). « Landscape agronomy : a new perspective for research on agricultural landscapes ». Dans : *Landscape Ecology*.

BONNEMAIRE, J (1995). *Pays, paysans, paysages: dans les Vosges du Sud : les pratiques agricoles et la transformation de l'espace*. Institut national de la recherche agronomique.

BÜHLER, E-A, A CAMARA, S LOPEZ-RIDAURA et C-T SOULARD (2010). « Farms and territories: crossing agronomy and geography to elaborate multifunctional farming systems ». Français. Dans : *Innovation and Sustainable Development in Agriculture and Food*. Sous la dir. d'E COUDEL, H DEVAUTOUR, C-T SOULARD et B HUBERT. Montpellier, France : Cirad-Inra-SupAgro, p. 16.

BUREL, F et J BAUDRY (2005). « Habitat quality and connectivity in agricultural landscapes: The role of land use systems at various scales in time ». Dans : *Ecological Indicators* 5.4, p. 305–313.

CASTELLAZZI, MS, GA WOOD, PJ BURGESS, J MORRIS, KF CONRAD et JN PERRY (2008). « A systematic representation of crop rotations ». Dans : *Agricultural Systems* 97.1-2, p. 26–33.

COPPEDGE, BR, DM ENGLE, RE MASTERS et MS GREGORY (2001). « Avian response to landscape change in fragmented southern Great Plains grasslands ». Dans : *Ecological Applications* 11.1, p. 47–59.

CREWS-MEYER, KA (2004). « Agricultural landscape change and stability in northeast Thailand: historical patch-level analysis ». Dans : *Agriculture, ecosystems & environment* 101.2-3, p. 155–169.

DEFFONTAINES, JP, C THENAIL et J BAUDRY (1995). « Agricultural systems and landscape patterns: how can we build a relationship? » Dans : *Landscape and urban planning* 31.1-3, p. 3–10.

DEFFONTAINES, JP et P THINON (2001). « Des entités spatiales significatives pour l'activité agricole et pour les enjeux environnementaux et paysagers. Contribution à une agronomie du territoire ». Dans : *Courrier de l'Environnement, Inra* 44, p. 13–28.

DOGLIOTTI, S, WAH ROSSING et MK VAN ITTERSUM (2003). « ROTAT, a tool for systematically generating crop rotations ». Dans : *European Journal of Agronomy* 19.2, p. 239–250.

DONALD, PF, RE GREEN et MF HEATH (2001). « Agricultural intensification and the collapse of Europe's farmland bird populations ». Dans : *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1462, p. 25.

FAOSTAT (2009). http://faostat.fao.org/site/377/default.aspx.

GLP (2005). *Science Plan and Implementation Strategy*. Rap. tech. IGBP Report No. 53/IHDP Report No. 19. IGBP Secretariat, Stockholm, p. 64.

HÉNIN, D (1999). *De la méthode en agronomie*. Ecologie et agronomie appliquées. l'Harmattan.

HÉNIN, S et JP DEFFONTAINES (1970). « Principe et utilité de l'étude des potentialités agricoles régionales ». Dans : *Comptes Rendus de l'Académie d'Agriculture de France*, p. 463–472.

HERRERA, JM et M LIEDGENS (2009). « Leaching and utilization of nitrogen during a spring wheat catch crop succession ». Dans : *Journal of environmental quality* 38.4, p. 1410–1419.

HOUET, T (2006). « Occupation des sols et gestion de l'eau: modélisation prospective en paysage agricole fragmenté (Application au SAGE du Blavet) ». Thèse de doct. Université Rennes 2.

INCHAUSTI, P et V BRETAGNOLLE (2005). « Predicting short-term extinction risk for the declining Little Bustard Tetrax tetrax in intensive agricultural habitats ». Dans : *Biological conservation* 122.3, p. 375–384.

INTERNATIONAL DE LA LANGUE FRANÇAISE, Conseil (1999). *Dictionnaire d'agriculture: français-anglais-allemand*. Le Conseil.

JACOPIN, R (2011). « Paysages et pratiques des agriculteurs dans le Sud du Plateau Lorrain : logiques d'organisation et effect sur l'environnement. » Thèse de doct. Université Nancy 2.

LACHIVER, M (1997). *Dictionnaire du monde rural: les mots du passé*. Fayard.

LANDAIS, E et JP DEFFONTAINES (1990). « Comprendre la gestion d'un espace pastoral. Étude monographique des pratiques d'un berger d'estive dans les Alpes du Sud ». Dans : *Recherches sur les Systèmes Herbagers*, p. 189–197.

LAZRAK, EG, M BENOÎT et J-F MARI (2009). « Fouille de données à l'aide de modèles stochastiques: segmentation temporo-spatiale des successsions de cultures d'un territoire agricole à l'aide de HMM2 ». Dans : *STIC 2009 Environnement*. Calais, France.

— (2010). « Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales ». Dans : *International Conference on Integrative Landscape Modelling*. Montpellier, France.

LAZRAK, EG, J-F MARI et M BENOÎT (2009). « Landscape regularity modelling for environmental challenges in agriculture ». Dans : *Landscape Ecology* 25.2, p. 169–183.

LAZRAK, EG, N SCHALLER et J-F MARI (2011). « Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol ». Français. Dans : *Atelier en marge d'EGC 2011*. Brest, France.

LE BER, F, M BENOÎT, C SCHOTT, J-F MARI et C MIGNOLET (2006). « Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software ». Dans : *Ecological Modelling* 191.1, p. 170–185.

LUNT, ID et PG SPOONER (2005). « Using historical ecology to understand patterns of biodiversity in fragmented agricultural landscapes ». Dans : *Journal of Biogeography* 32.11, p. 1859–1873.

MARI, J-F, EG LAZRAK et M BENOÎT (2010). « Fouille de paysages agricoles: analyse des voisinages des successions d'occupation du sol ». Dans : *Colloque RTE (Raisonnement sur le Temps et l'Espace) en marge de RFIA 2010*. Sous la dir. de F LE BER, G LIGOZAT, O PAPINI et M BOUZID. Caen, France.

MARI, J-F et F LE BER (2006). « Temporal and Spatial Data Mining with Second-Order Hidden Markov Models ». Dans : *Soft Computing*. ISSN:1432-7643 10.5, p. 406–414.

MARI, J-F, F LE BER et M BENOÎT (1998). « Reconnaissance de successions culturales par modèles de Markov : une étude préliminaire ». Dans : *Journées Cassini*. Marne-la-Vallée.

— (1999). « Classification de successions culturales par modèles de Markov ». Dans : *Septième journées de la Société Francophone de Classification - SFC'99*. Colloque avec actes et comité de lecture, p. 177–184.

MARTIN, P, F PAPY, V SOUCHÈRE et A CAPILLON (1998). « Maîtrise du ruissellement et modélisation des pratiques de production ». Dans : *Cahiers Agricultures* 7.2, p. 111–119.

MARTIN, P, A JOANNON, C MIGNOLET, V SOUCHÈRE et C THENAIL (2006). « Systèmes de culture et territoires: cas des questions environnementales ». Dans : *L'agronomie aujourd'hui*. Versailles : Quae éditions, p. 253–283.

MÉROT, A, C AUBRY, M BARBIER, A JOANNON, P MARTIN, C THENAIL et M BENOÎT (Soumis). « Interfacing landscape and agrosystems research at various scales to deal with natural resource preservation : a review ». Dans : *Agriculture Ecosystems & Environment*.

MIGNOLET, C, C SCHOTT et M BENOÎT (2004). « Spatial dynamics of agricultural practices on a basin territory: a retrospective study to implement models simulating nitrate flow. The case of the Seine basin ». Dans : *Agronomie* 24.4, p. 219–236.

— (2007). « Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale ». Dans : *Science of the Total Environment* 375.1-3, p. 13–32.

MOTTET, A, S LADET, N COQUE et A GIBON (2006). « Agricultural land-use change and its drivers in mountain landscapes: A case study in the Pyrenees ». Dans : *Agriculture, ecosystems & environment* 114.2-4, p. 296–310.

ORMEROD, SJ et AR WATKINSON (2000). « Editors' introduction: birds and agriculture ». Dans : *Journal of applied ecology* 37.5, p. 699–705.

OSTY, PL (1978). « L'exploitation vue comme une système: Diffusion de l'innovation et contribution au développement ». Dans : *Bulletin Technique d'Information* 326, p. 43–49.

PAPY, F et A TORRE (2002). « Quelles organisations territoriales pour concilier production agricole et gestion des ressources naturelles ». Dans : *Etudes et Recherches sur les Systèmes Agraires et le Développement* 33, p. 151–170.

POUDEVIGNE, I et D ALARD (1997). « Landscape and agricultural patterns in rural areas: a case study in the Brionne Basin, Normandy, France ». Dans : *Journal of Environmental Management* 50.4, p. 335–349.

RAINI, JA (2009). « Impact of land use changes on water resources and biodiversity of Lake Nakuru catchment basin, Kenya ». Dans : *African Journal of Ecology* 47, p. 39–45.

REY, A (2005). *Dictionnaire culturel en langue française: Coffret en 4 volumes*. Le Robert.

RINDFUSS, RR, SJ WALSH, BL TURNER, J FOX et V MISHRA (2004). « Developing a science of land change: Challenges and methodological issues ». Dans : *Proceedings of the National Academy of Sciences of the United States of America* 101.39, p. 13976–13981.

SCHALLER, N, EG LAZRAK, P MARTIN, J-F MARI, C AUBRY et M BENOÎT (2011). « Combining farmers' decision rules and landscape stochastic regularities for landscape modelling ». Anglais. Dans : *Landscape Ecology*.

SCHIETTECATTE, W, L D'HONDT, WM CORNELIS, ML ACOSTA, Z LEAL, N LAUWERS, Y ALMOZA, GR ALONSO, J DÍAZ, M RUÍZ et D GABRIELS (2008). « Influence of landuse on soil erosion risk in the Cuyaguateje watershed (Cuba) ». Dans : *CATENA* 74.1, p. 1–12.

SEBILLOTTE, M (1990). « Système de culture, un concept opératoire pour les agronomes ». Dans : *Les systèmes de culture*. Un Point sur. L Combe, p. 165–196.

— (2005). « Agronomes et territoires : Deuxième édition des Entretiens du Pradel ». Dans : sous la dir. de P PREVOST. Harmattan. Chap. Agronomes et territoires. Les trois métiers des agronomes, p. 479–497.

SEBILLOTTE, M et L-G SOLER (1988). « Le concept de modèle général et la compréhension du comportement de l'agriculteur, CR Acad ». Dans : *Agric. Fr* 74, p. 59–70.

— (1990). « Les processus de décision des agriculteurs ». Dans : *Modélisation systémique et système agraire: décision et organisation*. Sous la dir. de J BROSSIER, B VISSAC et JLL MOIGNE. INRA.

SIMPSON, JW, REJ BOERNER, MN DE MERS, LA BERNS, FJ ARTIGAS et A SILVA (1994). « Forty-eight years of landscape change on two contiguous Ohio landscapes ». Dans : *Landscape Ecology* 9.4, p. 261–270.

STOMPH, TJ, LO FRESCO et H VAN KEULEN (1994). « Land use system evaluation: Concepts and methodology ». Dans : *Agricultural systems* 44.3, p. 243–255.

THENAIL, C, A JOANNON, M CAPITAINE, V SOUCHÈRE, C MIGNOLET, N SCHERMANN, F DI PIETRO, Y PONS, C GAUCHEREL, V VIAUD et al. (2009). « The contribution of crop-rotation organization in farms to crop-mosaic patterning at local landscape scales ». Dans : *Agriculture, Ecosystems & Environment* 131.3-4, p. 207–219.

THINON, P (2003). « Les unités agro-physionomiques: quels usages? Quelle prise en compte du temps? » Dans : *Actes du colloque international*. Montpellier, France : P Dugué et Ph Jouve.

THINON, P et JP DEFFONTAINES (1999). « Partage de l'espace rural pour la gestion de problèmes environnementaux et paysagers dans le Vexin français ». Dans : *Cah Agr* 8, p. 373–87.

TSCHARNTKE, T, AM KLEIN, A KRUESS, I STEFFAN-DEWENTER et C THIES (2005). « Landscape perspectives on agricultural intensification and biodiversity-ecosystem service management ». Dans : *Ecology Letters* 8.8, p. 857–874.

TURNER, BL, EF LAMBIN et A REENBERG (2007). « The emergence of land change science for global environmental change and sustainability ». Dans : *Proceedings of the National Academy of Sciences* 104.52, p. 20666–20671.

TURNER, BL, D SKOLE, S SANDERSON, G FISCHER, L FRESCO et R LEEMANNS (1995). *Land-Use and Land-Cover Change: Science/Research Plan*. Rap. tech., p. 132.

VERBURG, PH et A VELDKAMP (2001). « The role of spatially explicit models in land-use change research: a case study for cropping patterns in China ». Dans : *Agriculture, ecosystems & environment* 85.1-3, p. 177–190.

VERBURG, PH, K KOK, RG PONTIUS et A VELDKAMP (2006). « Modeling Land-Use and Land-Cover Change ». Dans : *Land-Use and Land-Cover Change*. Sous la dir. d'EF LAMBIN et H GEIST. Global Change – The IGBP Series (closed). Springer Berlin Heidelberg, p. 117–135. ISBN : 978-3-540-32202-3.

VITOUSEK, PM, HA MOONEY, J LUBCHENCO et JM MELILLO (1997). « Human domination of Earth's ecosystems ». Dans : *Science* 277.5325, p. 494.

Chapitre 2

Corpus de données spatio-temporelles d'OCS et outils de fouille utilisés

Ce chapitre présente les contextes géographique et méthodologique de notre travail. Nous commençons par présenter les deux territoires d'étude et leurs corpus de données spatio-temporelles d'OCS, puis nous décrivons les outils de fouille de données utilisés pour la modélisation de l'OTAA.

2.1 Les territoires d'étude et leurs corpus de données spatio-temporelles d'OCS

2.1.1 Contexte géographique et problématique environnementale

Le site d'étude de Chizé : un territoire agricole de dimension moyenne

Le site d'étude de Chizé est situé dans la plaine de Niort au sud du département des Deux-Sèvres. C'est un territoire agricole de près de 450 km^2 qui a été labellisé zone atelier Plaine et Val de Sèvres en 2008 (figure 2.1). Depuis 1994, ce site fait l'objet de nombreuses études menées par le CNRS de Chizé en vue de comprendre, dans le cadre de programmes de recherche interdisciplinaires, les mécanismes impliqués dans le déclin de la biodiversité patrimoniale en plaine céréalière. Ces travaux concernent un ensemble d'espèces patrimoniales d'oiseaux de plaine (Busard cendré, Œdicnème criard, Outarde canepetière), leurs milieux de vie (la plaine agricole) et leurs espèces proies (insectes, passereaux et micromammifères).

Le bassin versant du Yar : un petit territoire régional

Le bassin versant du Yar est situé à l'extrême nord-ouest du département des Côtes-d'Armor, en limite du département du Finistère (figure 2.2). C'est un bassin versant littoral d'une superficie d'environ 60 km^2. Il est à l'origine de plus de 60 % des flux polluants alimentant la baie de Saint-Michel-en-Grève. Les cours d'eau du bassin versant du Yar apportent de grandes quantités d'éléments nutritifs sur le littoral et y entraînent un phénomène d'eutrophisation. Le bassin versant du Yar fait l'objet de cartographies d'OCS réalisées par le laboratoire Costel (Climat et Occupation du sol par télédétection) en vue de fournir les informations nécessaires à l'étude des facteurs responsables de cette eutrophisation (CORGNE, 2004).

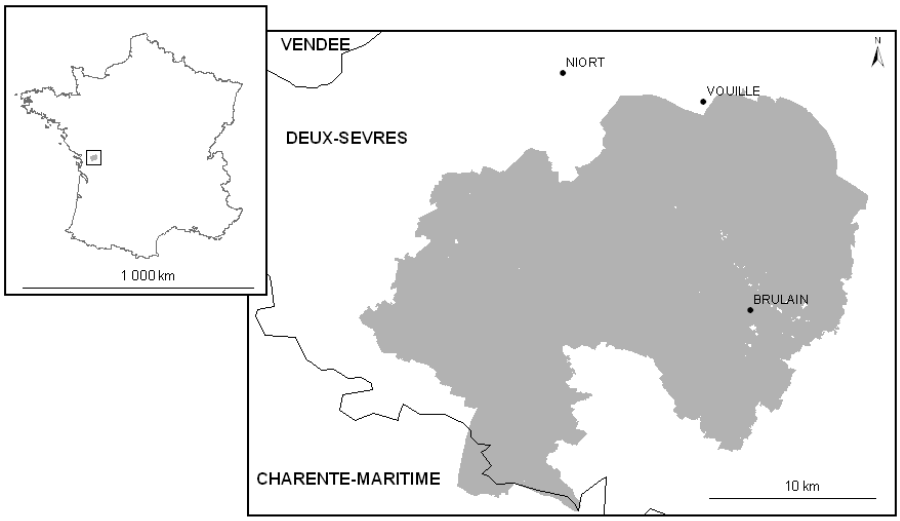


FIGURE 2.1 – Situation géographique de la zone atelier Plaine et Val de Sèvres.

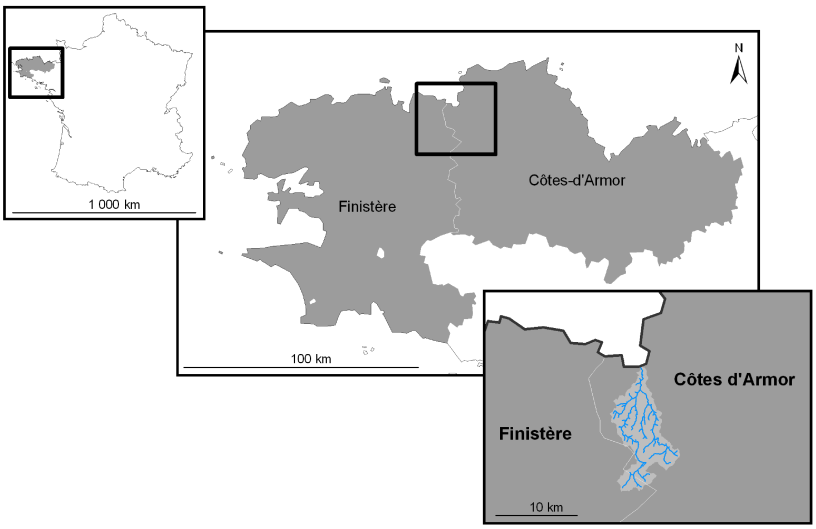


FIGURE 2.2 – Situation géographique du bassin versant du Yar.

2.1.2 Corpus de données d'OCS

Cas d'un corpus construit par des relevés annuels d'OCS : le site d'étude de Chizé

Dans le cadre de ses recherches sur la biodiversité en relation avec les pratiques agricoles, le Centre d'Études Biologiques de Chizé (CEBC) effectue, chaque année, depuis 1994, deux relevés d'OCS (avril et juin) sur la totalité de son site d'étude. Ces deux relevés annuels permettent de prendre en compte à la fois les cultures précocement récoltées et celles tardivement plantées. Le site d'étude de Chizé s'est progressivement élargi : de 20 km^2 en 1994 à 200 km^2 en 1995, puis 320 km^2 en 1996 avec une relative stabilisation jusqu'en 2005 suivie par d'autres élargissements à 420 km^2 en 2006 et à 430 km^2 en 2007. Dans ce travail de thèse, nous avons considéré la période de 12 années successives à partir de 1996 (figure 2.3). Lors des relevés d'OCS, les enquêteurs attribuent un code dans une nomenclature de 45 modalités d'OCS dont 40 sont agricoles (figure 2.4). Les frontières des parcelles étant susceptibles de changer chaque année en fonction des choix des agriculteurs, les enquêteurs définissent l'ensemble des **parcelles élémentaires** (ou micro parcelles) comme étant l'ensemble des polygones constitués par l'union des frontières des parcelles (figure 2.5). La zone d'étude compte environ 20 000 parcelles élémentaires, couvrant une superficie de 430 km^2. Les données collectées sont stockées dans un SIG au format vectoriel et constituent une couche d'informations spatio-temporelles où chaque parcelle élémentaire est liée à une succession d'OCS le long de la période d'étude.

Cas d'un corpus construit par traitement d'images satellites : le bassin versant du Yar

Pour le bassin versant du Yar, le laboratoire Costel a identifié les classes d'OCS par télédétection d'une série d'images satellites. Six OCS sont distinguées dont trois seulement sont agricoles. L'Urbain et la Forêt sont issus de l'analyse de deux images satellites, une en début de la période d'étude (1997), l'autre en son milieu (2003). Les autres OCS sont issues d'une identification annuelle à partir d'une série de 12 images prises en période estivale le long de la période d'étude : 1997 – 2008 (figure 2.6). Les données d'OCS issues du traitement des images satellites se présentent sous forme d'une couche vectorielle d'informations spatio-temporelles.

Comparaison entre les corpus des deux territoires d'étude

Le tableau 2.1 dresse une comparaison entre les deux territoires d'étude en mettant l'accent sur leurs caractéristiques contrastées. Dans ces deux cas d'étude, les bases de données spatio-temporelles des OCS sont construites et stockées sous SIG au format vectoriel. Dans le cas des données collectées par des relevés de terrain, les polygones correspondent à des parcelles élémentaires (*cf.* figure 2.5), alors que pour la base de données construite par traitement d'images satellites, les polygones correspondent à un ensemble de pixels voisins affectés à un même type d'OCS. Le nombre de modalités d'OCS peut aller d'une petite dizaine dans une base de données constituée par traitement d'images satellites à plusieurs dizaines dans une base de données constituée par des relevés de terrain.

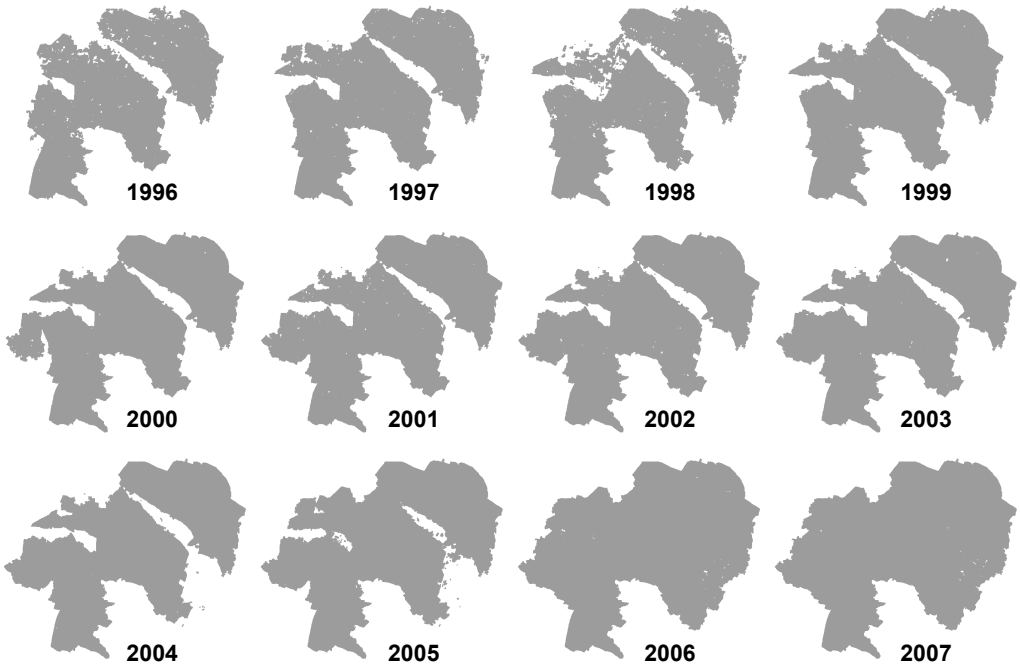


FIGURE 2.3 – Progression de la zone couverte par les relevés d'OCS pendant la période d'étude considérée dans ce travail de thèse (1996-2007).

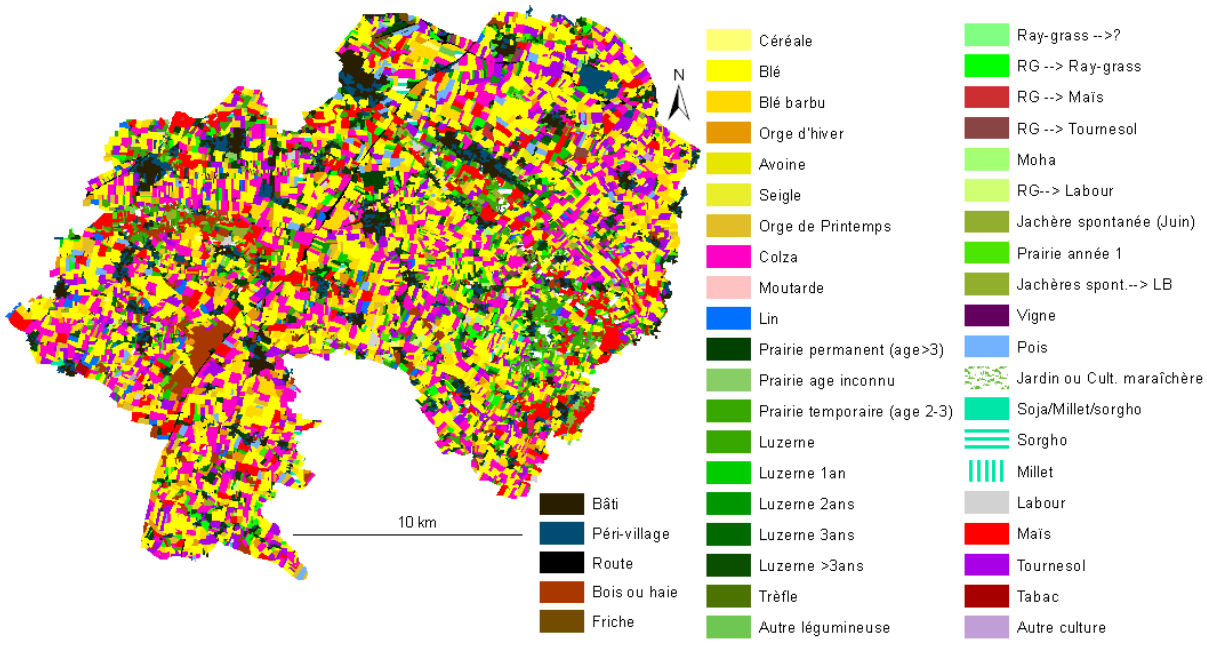


FIGURE 2.4 – Carte d'OCS du site d'étude de Chizé en 2007. La légende (à droite) compte 45 OCS dont 40 agricoles, 3 urbaines (Bâti, Péri village, Route) et deux forestières (Bois ou haie, Friche).

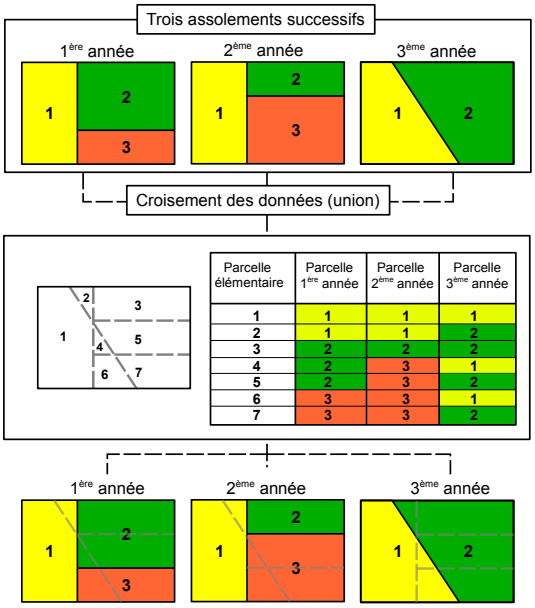


FIGURE 2.5 – Évolution des limites parcellaires sur trois années consécutives. Dans cet exemple, l'union des limites parcellaires durant cette période conduit à la définition de sept parcelles élémentaires. Chaque parcelle élémentaire a hébergé une succession de cultures. Par exemple, la parcelle élémentaire n° 4 a hébergé la succession : 2-3-1.

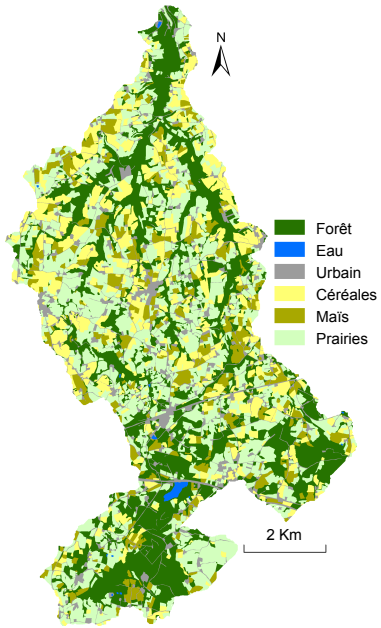


FIGURE 2.6 – Carte d'OCS du bassin versant du Yar en 2008. La légende compte 6 OCS dont 3 sont agricoles.

TABLEAU 2.1 – Comparaison entre les deux bases de données d'OCS utilisées.

	Site d'étude de Chizé	Bassin versant du Yar
Source de données	Relevés de terrain	Images satellites
Surface (km^2)	~ 350	~ 60
Période d'étude	1996–2007 [a]	1997 – 2008
Nombre de modalités	45	6
Représentation spatiale	Vectorielle	Matricielle (convertie en vectorielle)
Entités spatiales élémentaires	Parcelles élémentaires (polygones)	Pixels ($20 \times 20\,m^2$)
Problématique environnementale	Biodiversité (oiseaux patrimoniaux)	Qualité de l'eau (eutrophisation)
Format de la base de données	Shapefile [b]	Shapefile

a. Les relevés de terrain se poursuivent dans la zone atelier Plaine et Val de Sèvre, mais pour que les résultats obtenus dans le cadre de cette thèse gardent leur cohérence, nous avons choisi d'utiliser la même période d'étude dans l'ensemble de ce travail.

b. Format populaire de données vectorielles géospatiales affichable dans les logiciels des systèmes d'information géographiques.

2.2 Fouille de données spatio-temporelles d'OCS par segmentation à l'aide de modèles de Markov cachés

2.2.1 Fouille de données et recherche de régularités : définitions

Nous désignons par « ***fouille de données*** » la classification de ces données en vue d'interprétation par un ***expert du domaine*** qui possède, sous forme de ***règles***, une certaine connaissance du domaine (p.ex. agronomie, écologie, hydrologie, ...) et du territoire d'étude. Nous utilisons les modèles de Markov cachés d'ordre 2 HMM2 pour fouiller les données spatio-temporelles d'OCS moyennant une classification probabiliste où les classes sont représentées par des distributions d'observations. Cette classification permet de faire émerger des structures inconnues et potentiellement utiles : ***les régularités***. L'expert du domaine d'étude interprète ces régularités qui lui permettent d'améliorer sa connaissance du domaine et d'interagir avec l'***analyste*** [1] pour formuler d'autres besoins de fouille de données (figure 2.7).

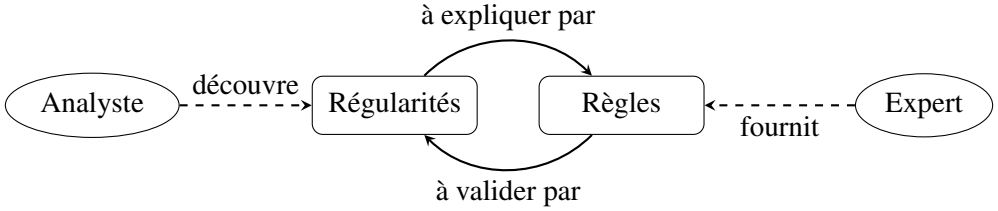


FIGURE 2.7 – Articulation des régularités issues de la fouille de données avec les connaissances sous forme de règles logiques fournies par un expert du domaine.

Nous commençons cette section par une définition des modèles de Markov cachés suivie de la présentation de deux de leurs propriétés en rapport avec notre approche de modélisation de l'OTAA : l'apprentissage automatique et la segmentation. Nous illustrons ces propriétés à l'aide d'un exemple pédagogique

1. ***L'analyste*** est celui qui réalise la fouille de données.

sur les HMM2. Nous terminons cette section par la description de la boîte à outils « ARPENTAGE » qui implémente la procédure de fouille mise en œuvre dans le cadre de cette thèse.

2.2.2 Définition d'un HMM

Les modèles de Markov cachés (HMM) sont issus du domaine de la reconnaissance de la parole (JELINEK, 1976) et de la reconnaissance des formes (BENMILOUD et PIECZYNSKI, 1995). Ils sont devenus des outils pertinents pour l'analyse de données séquentielles comme le génome en bioinformatique (KROGH, 1998) et les successions de cultures en agronomie (LE BER et al., 2006).

Un HMM est défini par la donnée d'un ensemble E, et de deux matrices A et B définis comme suit :

- $E = \{e_1, e_2, \ldots, e_N\}$, un ensemble fini à N états.
- A est la matrice définissant les probabilités des transitions entre les états du modèle.
- B est la matrice regroupant les distributions des observations $b_i(\cdot)$ associées aux états e_i. Ces distributions peuvent être paramétriques, non paramétriques ou bien être données par un HMM lorsqu'il s'agit d'un HMM hiérarchique (HHMM) (FINE et al., 1998 ; MARI et LE BER, 2006).

La matrice A définit la topologie du graphe des transitions entre états. Elle permet de spécifier quelles sont les transitions autorisées. Les transitions entre les états dépendent, suivant l'ordre n du modèle, de l'état courant et des n états précédents. Dans un HMM d'ordre 1 (HMM1), l'état suivant (e_j) dépend uniquement de l'état courant (e_i). La matrice de transitions d'un HMM1 est définie sur E^2 par $A = (aij)$. Dans un HMM d'ordre 2 (HMM2), l'état suivant (e_k) dépend de l'état courant (e_j) et de l'état précédent (e_i). La matrice de transitions ($A = (aijk)$) d'un HMM2 est tridimensionnelle (définie sur E^3). Comparés aux HMM1, les HMM2 permettent une meilleure modélisation des phénomènes transitoires (MARI et al., 1997).

Contrairement à la chaîne de Markov (CASTELLAZZI et al., 2008) où chaque état est associé à une observation unique (on parle d'états observés), dans un HMM, les observations sont le résultat d'une variable aléatoire caractérisée par la densité $b_i(\cdot)$ qui dépend de l'état e_i.

Un HMM modélise deux processus stochastiques :

Le premier est caché de l'observateur. Il est défini sur un ensemble d'états et correspond à une chaîne de Markov ;

Le second est visible. Il produit une observation à chaque pas de temps et dépend de la fonction de densité de probabilité associée à l'état dans le quel la chaîne de Markov se trouve au temps t.

La figure 2.8 donne un exemple d'un HMM1 à 3 états. Cet HMM1 est défini par sa matrice de transitions A et sa matrice B regroupant les distributions des observations associées aux états :

$$A = \begin{pmatrix} 0{,}3 & 0{,}5 & 0{,}2 \\ 0 & 0{,}3 & 0{,}7 \\ 0 & 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \\ 0.5 & 0{,}5 \\ 0 & 1 \end{pmatrix}$$

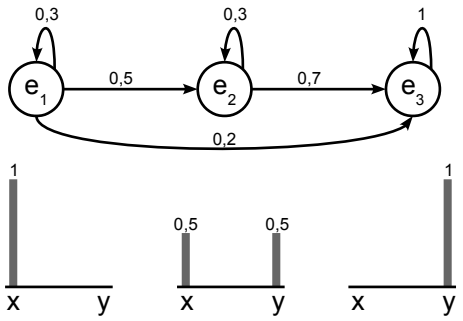


FIGURE 2.8 – Exemple d'un HMM1 avec 3 états et 2 observations notées x et y. Sous chaque état est représentée la distribution des 2 observations sous forme d'un diagramme en bâtons.

2.2.3 Apprentissage automatique d'un HMM

Les HMM sont des modèles dont les paramètres (A et B) peuvent être estimés par apprentissage automatique à partir d'un corpus d'observations et d'un modèle initial. L'apprentissage des HMM est généralement réalisé avec l'algorithme Forward-Backward – également appelé Baum-Welch – qui est un algorithme du type EM (BAUM et al., 1970 ; WELCH, 2003). L'estimation est un processus itératif qui commence par un modèle initial et un corpus défini par une suite d'observations auxquelles le HMM doit s'adapter. Habituellement, le modèle est initialisé avec un même poids des probabilités de transitions et une distribution uniforme dans chaque état. A chaque itération, l'algorithme de Forward-Backward détermine un nouveau modèle dans lequel la vraisemblance de la suite d'observations augmente en convergeant vers un optimum local qui dépend du modèle initial. La vraisemblance, notée $L(o)$, est définie par la somme des probabilités jointes de tous les alignements « suite d'états Q_1^T, suite d'observations $O_1^T = o$ ».

$$L(o) = \sum_{Q} P(Q, O_1^T = o) \tag{2.1}$$

Lorsque N est le nombre d'états et T la longueur de la séquence observée, l'algorithme Forward-Backward a une complexité de N^2T pour un HMM1 et de N^3T pour un HMM2.

Les HMM sont des outils efficaces de fouille de données et de segmentation spatio-temporelle. Ils présentent l'avantage de posséder des algorithmes simples et rapides pour l'estimation de leurs paramètres comparés aux algorithmes d'apprentissage des champs de Markov cachés. Nous présentons dans ce qui suit, le logiciel ARPENTAGE qui est une boîte à outils de fouille de données fondée sur les HMM d'ordre 2.

2.2.4 Segmentation à l'aide d'un HMM2

La segmentation avec un HMM2 consiste à affecter les observations du corpus à l'un des états effectifs du modèle. Nous utilisons l'algorithme Classification EM (CEM) mis au point par CELEUX et GOVAERT (1992). Cet algorithme définit des segments temporels ou des zones homogènes correspondant aux états

du HMM2 caractérisés par des distributions probabilistes des observations. Cet algorithme se déroule en 3 étapes :

1. calcul des probabilités *a posteriori* de tous les alignements possibles ;
2. utilisation du critère MAP (Maximum *a posteriori* probability) qui consiste à affecter chaque observation o_t de la suite d'observations O_1^T à l'état qui maximise la probabilité *a posteriori* : $P_t(e_i|O_1^T = o)$;
3. mise à jour des paramètres du HMM2 en maximisant la vraisemblance.

2.2.5 Approximation d'un champ de Markov caché par un HMM

Motivation

À des échelles régionales, nous considérons que l'espace agricole peut être vu comme une mosaïque de polygones sur lesquelles les OCS se succèdent à un rythme annuel ou pluriannuel. Par analogie avec une image numérique où le niveau de gris constitue l'attribut de chaque pixel, dans notre représentation de l'espace agricole, les parcelles sont des pixels de formes et de tailles irrégulières et les successions d'OCS sont les attributs de ces pixels. Nous faisons l'hypothèse que la succession d'OCS sur une parcelle dépend des successions du voisinage (temporel et spatial) proche définissant un champ de Markov de successions que nous analysons en empruntant et en adaptant des techniques d'analyse propres au traitement d'image.

Les champs de Markov cachés sont des outils de fouille de données permettant la classification de données spatialisées assimilables à un champ de Markov. Dans ces modèles, les relations d'interdépendance spatiale sont décrites par un graphe non orienté où la dépendance spatiale ne concerne que les voisins immédiats (propriété markovienne). Les champs de Markov cachés sont depuis longtemps utilisés pour résoudre des problèmes de classification spatiale comme en classification d'images et plus généralement en reconnaissance de formes (BERTHOD et al., 1996). Les champs de Markov cachés sont également utilisés en cartographie vue comme une classification automatique de données spatialisées. À titre d'exemple, AZIZI et al. (2011) utilisent des champs de Markov cachés pour cartographier le risque de maladies rares (AZIZI et al., 2011). Mais en raison de la complexité de leurs algorithmes d'estimation, les champs de Markov cachés sont longtemps restés peu utilisés pour la classification de données spatio-temporelles (ABRIAL et al., 2010).

Technique

La transformation d'une image bidimensionnelle en une séquence monodimensionnelle par un parcours de type Hilbert-Peano (figure 2.9) rend possible l'utilisation des modèles de Markov cachés caractérisés par des algorithmes d'apprentissage moins complexes et plus rapides pour résoudre des problèmes de segmentation statistique des données spatio-temporelles (BENMILOUD et PIECZYNSKI, 1995). L'introduction d'un tel parcours est particulièrement aisée dans le cas des images carrées de taille $2^n \times 2^n$. Ces auteurs montrent que les HMM donnent des performances comparables aux champs de Markov cachés.

Adaptation au contexte agricole

Pour modéliser l'OTAA que nous représentons par un champ de Markov de successions d'OCS, nous suivons la technique de segmentation proposée par BENMILOUD et PIECZYNSKI (1995) que nous adaptons aux particularités que présente l'OTAA par rapport à une image carrée constituée d'une matrice carrée de pixels :

Particularité 1 : les polygones formant la mosaïque agricole sont de taille et de forme quelconques Afin de traiter le problème de la forme irrégulière des parcelles, la couche d'information vectorielle relative au territoire d'étude est rastérisée en échantillonnant les attributs des polygones avec une grille carrée de $2^n \times 2^n$ points régulièrement espacés, et l'espace 2D agricole est transformé en une séquence monodimensionnelle par le parcours de Peano des points de la grille d'échantillonnage. Le corpus résultant est une matrice où les colonnes représentent les OCS année par année et les lignes, les différents points de la grille ordonnés par le parcours de Peano (figures 2.10 (a) et (b)).

Particularité 2 : Les polygones de la mosaïque agricole définissent un système de voisinage irrégulier L'analyse des voisinages des parcelles avec des chaînes de Markov cachées nécessite de trouver un chemin pour parcourir l'ensemble des parcelles du territoire étudié tout en préservant l'information du voisinage. Définir un tel chemin est un problème qui ne possède pas d'algorithme pouvant le résoudre dans un temps raisonnable, c'est un problème dit Non Polynomial (NP) difficile. La solution approchée que nous avons adoptée pour traiter ce problème consiste à utiliser le parcours de la courbe fractale de Peano et à ajuster la profondeur de la fractale à la taille des polygones. Cet ajustement est réalisé en réduisant, à son centre de gravité, chaque motif de la fractale se trouvant à l'intérieur d'une parcelle élémentaire (figures 2.10 (c)).

Particularité 3 : Les territoires d'étude sont de forme quelconque Les territoires étudiés n'étant pas carrés, un grand nombre de points de la grille carrée d'échantillonnage se retrouve en dehors de la zone d'étude. Ce nombre de points doit être réduit pour ne pas perturber les résultats de la classification. Afin de traiter ce problème, les motifs de la fractale de Peano se trouvant en dehors de la zone d'étude sont réduits à un point à leur centre de gravité (figure 2.11). Le nombre de polygones ajoutés pour donner au territoire d'étude une forme carrée est considérablement réduit.

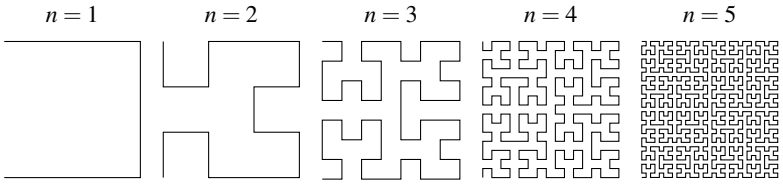


FIGURE 2.9 – Exemples de parcours de type Hilbert-Peano de tailles ($2^n \times 2^n$) croissantes.

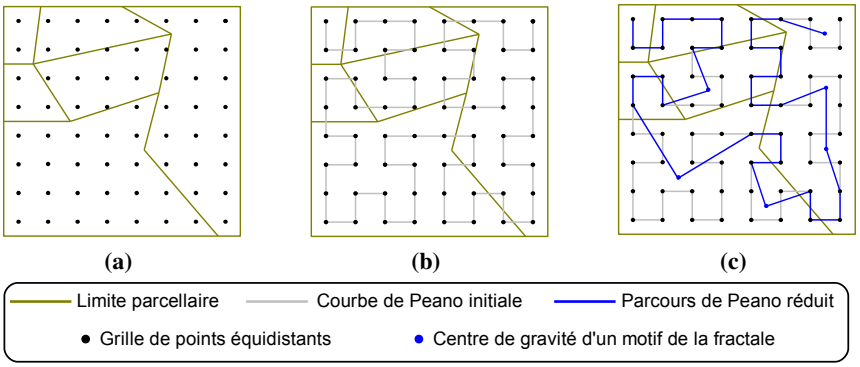


FIGURE 2.10 – Transformation du plan (2D) en séquence (1D). (a) Mosaïque de parcelles rastérisée par échantillonnage avec une grille de points équidistants. (b) Parcours de la grille de points avec une courbe fractale de Peano. Deux points voisins sur la courbe sont voisins dans l'espace mais l'inverse n'est pas vrai. (c) Ajustement de la profondeur de la courbe fractale de Peano à la taille des polygones.

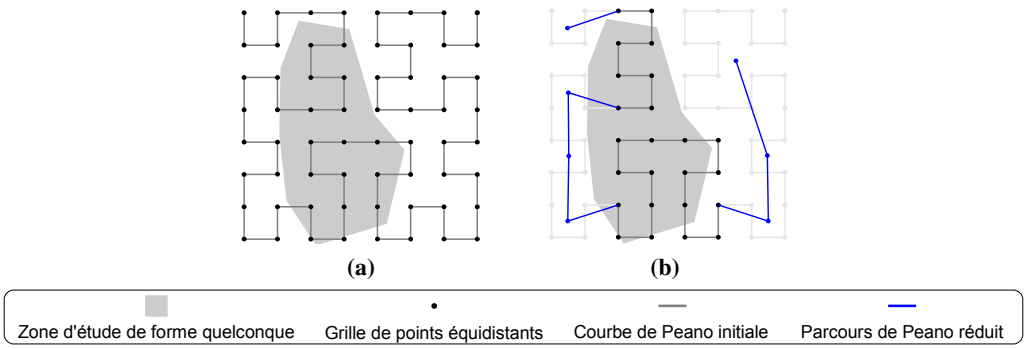


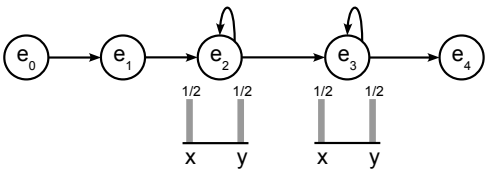
FIGURE 2.11 – Réduction des points de la fractale se trouvant en dehors de la zone d'étude. (a) Zone d'étude de forme quelconque dans la grille carrée des points équidistants parcourus par la fractale de Peano. Le nombre important de points en dehors de la zone d'étude risque de noyer l'information. (b) Les motifs de la fractale se trouvant entièrement en dehors de la zone d'étude sont réduits à un point à leur centre de gravité. Il en résulte une réduction significative des points en dehors de la zone d'étude et une amélioration de la segmentation par les HMM2.

2.2.6 Exemple d'une segmentation avec un HMM2

Cet exemple vise à illustrer les principes de l'apprentissage et de la segmentation pour les HMM. Un tutoriel très complet sur les HMM1 peut être trouvé dans RABINER (1989). Nous nous intéressons au cas particulier des HMM2 implémentés dans ARPENTAGE.

Considérons le HMM2 décrit dans la figure 2.12. Les états e_2 et e_3 sont associés chacun à une distribution uniforme de deux observations : x et y, et les transitions définies entre les états sont initialisées avec le même poids. Les probabilités de transitions entre les états sont calculées, à partir des poids des transitions, pour que la somme des probabilités de transitions en sortie de l'état e_j sachant l'état précédent e_i soit égale à 1 :

$$\forall(e_i, e_j) : \text{deux états successifs}, \sum_l a_{ijl} = 1 \tag{2.2}$$



(a)

Transition	Poids	a_{ijk}
$e_0e_1e_2$	1	1
$e_1e_2e_2$	1	1/2
$e_1e_2e_3$	1	1/2
$e_2e_2e_2$	1	1/2
$e_2e_2e_3$	1	1/2
$e_2e_3e_3$	1	1/2
$e_2e_3e_4$	1	1/2
$e_3e_3e_3$	1	1/2
$e_3e_3e_4$	1	1/2

(b)

FIGURE 2.12 – Exemple d'un HMM2 initial caractérisé par (a) 2 états effectifs (e_2 et e_3) et 2 observations notées x et y. Les deux états initiaux (e_0 et e_1) et l'état final (e_4) ne sont pas associés à des observations. Sous chaque état effectif est représentée la distribution des 2 observations sous forme d'un diagramme en bâtons. Les distributions des observations dans les états effectifs sont équiprobables. (b) Les transitions sont initialisées avec le même poids.

L'utilisation d'un HMM2 pour la segmentation d'une suite d'observations passe par l'estimation de ses paramètres. Comme nous l'avons vu en section 2.2.3 page 24, les distributions d'observations ($b_i(\cdot)$) associées à chaque état e_i et les probabilités de transitions (a_{ijk}) sont estimables à partir d'un corpus de données défini par une suite d'observations. Décrivons le principe d'estimation de ces paramètres sur le HMM2 de l'exemple (figure 2.12) en utilisant un corpus défini par une suite de 4 observations : $x-x-y-y$. L'estimation étant un processus itératif, nous détaillerons le calcul pour la première itération.

Un alignement est une suite d'états permettant d'observer la suite d'observations définissant le corpus. Dans cet exemple, il y a 3 alignements possibles. La figure 2.13 donne ces alignements en assimilant

les états à des urnes destinées à contenir les observations de la suite $x-x-y-y$. L'estimation des paramètres du HMM2 nécessite le calcul des probabilités de ces alignements.

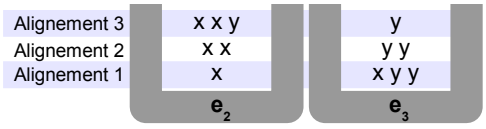


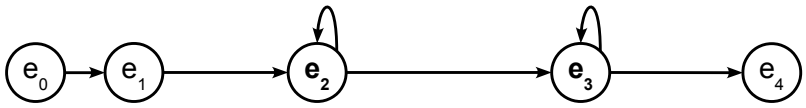
FIGURE 2.13 – Les trois alignements possibles permettant d'observer la suite $x-x-y-y$ avec le HMM2 de l'exemple.

Calcul des probabilités d'alignements

L'objectif de cette étape est de calculer la part de contribution de chaque alignement à l'ensemble de tous les alignements. Ces contributions relatives des alignements nous serviront pour l'estimation des paramètres du modèle. Elles sont recalculées à chaque itération. La contribution d'un alignement l est calculée par le rapport de la probabilité de cet alignement sur la somme des probabilités de tous les alignements :

$$Contrib(alignement\,l) = \frac{P(alignement\,l)}{\sum_{l} P(alignement\,l)} \tag{2.3}$$

Avant de calculer les probabilités des alignements, il faut d'abord identifier les transitions impliquées dans chaque alignement. La figure 2.14 illustre la procédure d'obtention des transitions mises en œuvre dans les 3 alignements de notre exemple (*cf.* figure 2.13).



Alignement 1		**x**		**x**	**y**	**y**
Transitions pour obtenir l'alignement 1		$e_0e_1e_2$		$e_1e_2e_3$	$e_2e_3e_3$	$e_3e_3e_3$ $e_3e_3e_4$
Alignement 2		**x**	**x**		**y**	**y**
Transitions pour obtenir l'alignement 2		$e_0e_1e_2$	$e_1e_2e_2$		$e_2e_2e_3$	$e_2e_3e_3$ $e_3e_3e_4$
Alignement 3		**x**	**x**	**y**	**y**	
Transitions pour obtenir l'alignement 3		$e_0e_1e_2$	$e_1e_2e_2$	$e_2e_2e_2$	$e_2e_2e_3$	$e_2e_3e_4$

FIGURE 2.14 – Inventaire des transitions permettant d'obtenir tous les alignements. Par exemple, pour l'alignement 1, la transition $e_0e_1e_2$ permet d'observer x dans e_2 et les transitions $e_1e_2e_3$, $e_2e_3e_3$, $e_3e_3e_3$ permettent respectivement d'observer x, y et y dans e_3. Enfin, la transition $e_3e_3e_4$ termine sans émettre d'observation.

Calculons maintenant les probabilités des alignements permettant d'observer la suite $x-x-y-y$ avec le HMM2 de l'exemple :

$$\begin{aligned} P(alignement\,1) &= a_{012}b_2(x)a_{123}b_3(x)a_{233}b_3(y)a_{333}b_3(y)a_{334} \\ &= 1\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2) \\ &= 1/2^8 \end{aligned}$$

$$\begin{aligned} P(alignement\,2) &= a_{012}b_2(x)a_{122}b_2(x)a_{223}b_3(y)a_{233}b_3(y)a_{334} \\ &= 1\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2) \\ &= 1/2^8 \end{aligned}$$

$$\begin{aligned} P(alignement\,3) &= a_{012}b_2(x)a_{122}b_2(x)a_{222}b_2(y)a_{223}b_3(y)a_{234} \\ &= 1\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2)\cdot(1/2) \\ &= 1/2^8 \end{aligned}$$

Nous pouvons maintenant calculer les contributions de chaque alignement à l'ensemble des alignements. Pour la première itération de notre exemple, les contributions des trois alignements sont égales :

$$Contrib(alignement\,1) = Contrib(alignement\,2) = Contrib(alignement\,3) = \frac{1}{3} \tag{2.4}$$

Principe d'estimation des probabilités de transitions entre les états

L'estimation des probabilités de transitions s'appuie sur le calcul des comptes (C_{ijk}) des transitions utilisées pour obtenir l'ensemble des alignements. Pour chaque transition $e_i e_j e_k$, C_{ijk} est calculé comme suit (équation 2.5) :

$$C_{ijk} = \sum_l (N_{ijk}(align.l) \times Contrib(align.l)) \tag{2.5}$$

N_{ijk} est le nombre d'utilisations de la transition $e_i e_j e_k$ dans un alignement donné.

$Contrib(align.l)$ est la contribution relative de cet alignement par rapport à l'ensemble des alignements (*cf.* équation 2.3).

Lors de l'apprentissage du HMM2 sur le corpus de données, les probabilités de transitions sont recalculées à chaque itération à partir des probabilités *a posteriori* (équation 2.6) :

$$a_{ijk} = \frac{C_{ijk}}{\sum_{l} C_{ijl}} \tag{2.6}$$

Pour notre exemple, les étapes de calcul des probabilités de transitions après une itération sont données par les tableaux de la figure 2.15.

(a) Calcul des comptes pour l'alignement 1

Transition	$N_{ijk} \times Contrib(align.1)$
$e_0e_1e_2$	$1 \times 1/3$
$e_1e_2e_3$	$1 \times 1/3$
$e_2e_3e_3$	$1 \times 1/3$
$e_3e_3e_3$	$1 \times 1/3$
$e_3e_3e_4$	$1 \times 1/3$

(b) Calcul des comptes pour l'alignement 2

Transition	$N_{ijk} \times Contrib(align.2)$
$e_0e_1e_2$	$1 \times 1/3$
$e_1e_2e_2$	$1 \times 1/3$
$e_2e_2e_3$	$1 \times 1/3$
$e_2e_3e_3$	$1 \times 1/3$
$e_3e_3e_4$	$1 \times 1/3$

(c) Calcul des comptes pour l'alignement 3

Transition	$N_{ijk} \times Contrib(align.3)$
$e_0e_1e_2$	$1 \times 1/3$
$e_1e_2e_2$	$1 \times 1/3$
$e_2e_2e_2$	$1 \times 1/3$
$e_2e_2e_3$	$1 \times 1/3$
$e_2e_3e_4$	$1 \times 1/3$

(d) Calcul des comptes de l'ensemble des alignements et estimation des probabilités de transitions

Transition	Comptes (C_{ijk})	Probabilités de transitions (a_{ijk})
$e_0e_1e_2$	3/3	1
$e_1e_2e_2$	2/3	2/3
$e_1e_2e_3$	1/3	1/3
$e_2e_2e_2$	1/3	1/3
$e_2e_2e_3$	2/3	2/3
$e_2e_3e_3$	2/3	2/3
$e_2e_3e_4$	1/3	1/3
$e_3e_3e_3$	1/3	1/3
$e_3e_3e_4$	2/3	2/3

FIGURE 2.15 – Estimation des probabilités de transitions du HMM2 de l'exemple (figure 2.12) à partir de la suite d'observations $x-x-y-y$.

Principe d'estimation des densités de probabilités associées aux états

Pour estimer les probabilités des observations associées aux états e_2 et e_3, le compte des observations émises par chaque état est calculé pour chaque alignement puis normalisé en probabilités d'émission d'observations $b_i(\cdot)$. Schématisons chaque état associé aux observations par une urne destinée à contenir les observations émises par les états du HMM2 pour les différents alignements. La figure 2.16 illustre le remplissage des urnes et donne les valeurs estimées des nouvelles distributions d'observations suite à l'apprentissage du HMM2 sur la suite d'observations $x-x-y-y$.

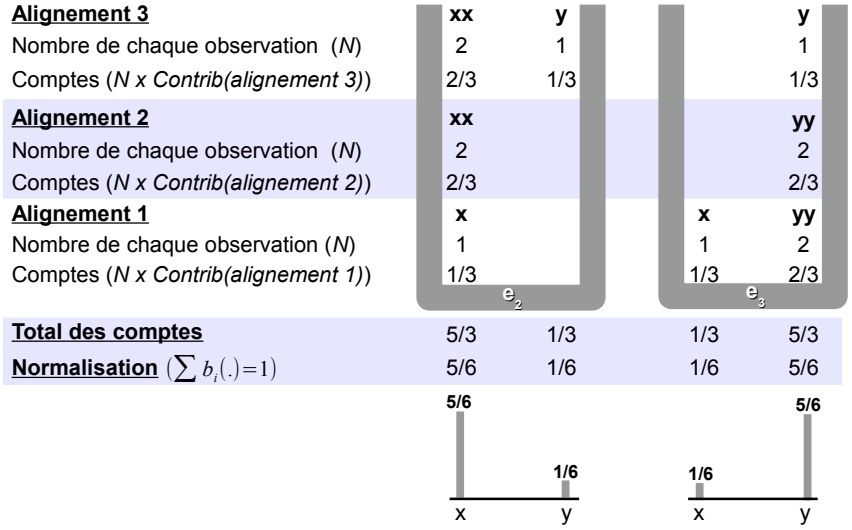
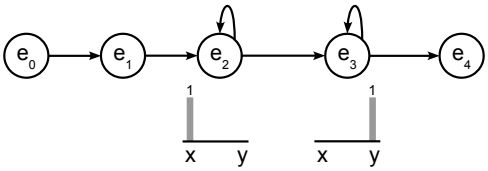


FIGURE 2.16 – Estimation des densités d'observations associées aux états effectifs (e_2 et e_3) du HMM2 de la figure 2.12.

Itérations et convergence vers un optimum local

Nous venons de dérouler le principe de la première itération de l'algorithme Forward-Backward (*cf.* section 2.2.3 page 24) pour l'estimation des paramètres ($b_i(\cdot)$ et a_{ijk}) du HMM2 de l'exemple. L'itération suivante suit le même principe en remplaçant les probabilités de transition du tableau 2.12(b) par celles du tableau (d) de la figure 2.15, et les distributions des observations initiales (figure 2.12(a)) par celles de la figure 2.16. Pour cet exemple simple où le modèle s'adapte parfaitement au corpus de données représenté par la suite $x-x-y-y$, l'estimation converge vers l'optimum dès la troisième itération (figure 2.17).



(a)

Transition	a_{ijk}
$e_0e_1e_2$	1
$e_1e_2e_2$	1
$e_1e_2e_3$	0
$e_2e_2e_2$	0
$e_2e_2e_3$	1
$e_2e_3e_3$	1
$e_2e_3e_4$	0
$e_3e_3e_3$	0
$e_3e_3e_4$	1

(b)

FIGURE 2.17 – HMM2 résultant de l'apprentissage du modèle de la figure 2.12(a) sur le corpus défini par la suite de 4 observations : $x-x-y-y$. Les états effectifs sont associés chacun à une seule observation. e_2 est associé à x et e_3 à y. Après l'apprentissage du modèle de l'exemple, les états ne sont plus « cachés » ils sont dits états de Dirac.

Segmentation de la suite d'observations

La segmentation de la suite d'observations $x-x-y-y$ avec le HMM2 appris de notre exemple consiste à affecter les observations qui se « ressemblent » à un état selon un critère de segmentation. Notre critère de segmentation (MAP, *cf.* section 2.2.4, page 24) consiste à parcourir la suite d'observations et à affecter à chaque observation l'état qui maximise la probabilité *a posteriori* [2]. Cette probabilité est calculée au temps t pour chaque état e_i comme suit :

$$P_t(e_i|O_1^T = o) = \frac{\sum P(\text{alignements qui passent par } e_i)}{\sum P(\text{chaque alignement})} \tag{2.7}$$

Nous illustrons le principe de la segmentation à l'aide du HMM2 de l'exemple appris avec la première itération. Rappelons que ce modèle est caractérisé par sa matrice de transition donnée dans le tableau(d) de la figure 2.15 et par les densités de probabilités associées aux états données dans la figure 2.16. Pour calculer les probabilités *a posteriori* nous aurons besoin de calculer les probabilités de chaque alignement (équations 2.8, 2.9 et 2.10). La figure 2.18 illustre le principe de cette segmentation.

$$\begin{aligned} P(alignement\,1) &= a_{012}b_2(x)a_{123}b_3(x)a_{233}b_3(y)a_{333}b_3(y)a_{334} \\ &= 1\cdot(5/6)\cdot(1/3)\cdot(1/6)\cdot(2/3)\cdot(5/6)\cdot(1/3)\cdot(5/6)\cdot(2/3) \\ &= 0{,}005 \end{aligned} \tag{2.8}$$

$$\begin{aligned} P(alignement\,2) &= a_{012}b_2(x)a_{122}b_2(x)a_{223}b_3(y)a_{233}b_3(y)a_{334} \\ &= 1\cdot(5/6)\cdot(2/3)\cdot(5/6)\cdot(2/3)\cdot(5/6)\cdot(2/3)\cdot(5/6)\cdot(2/3) \\ &= 0{,}05 \end{aligned} \tag{2.9}$$

$$\begin{aligned} P(alignement\,3) &= a_{012}b_2(x)a_{122}b_2(x)a_{222}b_2(y)a_{223}b_3(y)a_{234} \\ &= 1\cdot(5/6)\cdot(2/3)\cdot(5/6)\cdot(1/3)\cdot(1/6)\cdot(2/3)\cdot(5/6)\cdot(1/3) \\ &= 0{,}005 \end{aligned} \tag{2.10}$$

Si $x-x-y-y$ est une suite d'observations décrivant un parcours de Peano dans le plan (figure 2.19 (a)), la segmentation de cette suite d'observations définit deux classes spatiales : e_2 et e_3 (figure 2.19 (b)). Ces

2. Le terme ***a posteriori*** vient du fait que la probabilité est calculée à partir des comptes, c.à.d après avoir parcouru toute la suite d'observations.

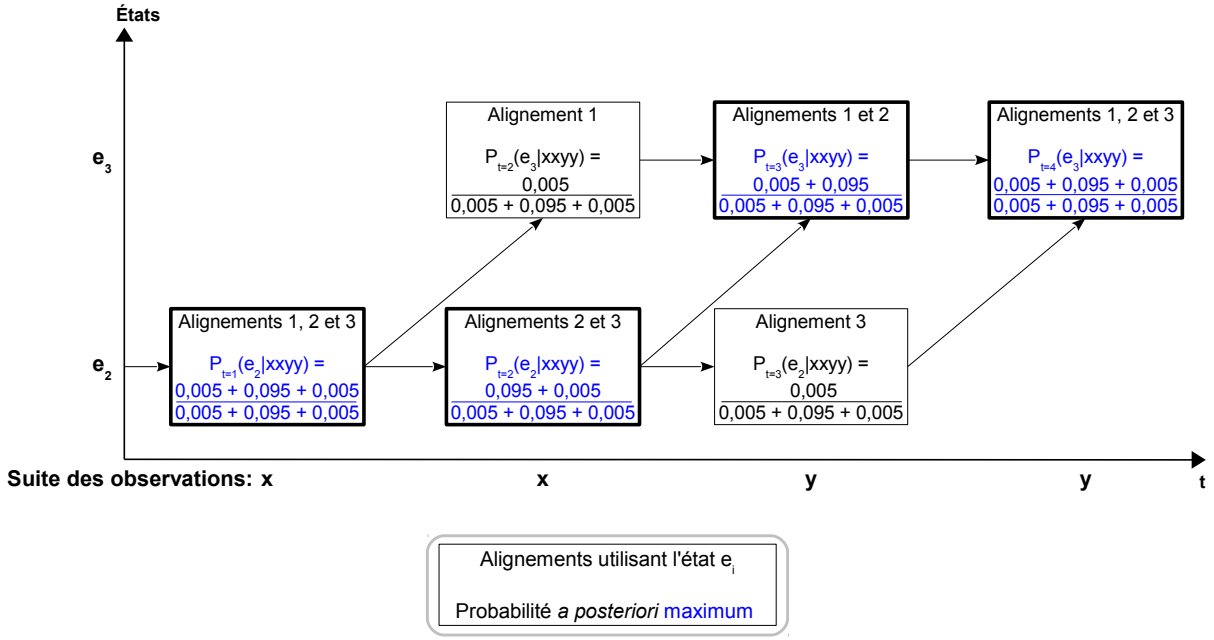


FIGURE 2.18 – Segmentation de la suite $x-x-y-y$ avec le HMM2 de l'exemple appris avec une itération. Deux classes sont identifiées : la classe représentée par l'état e_2 contient le segment $x-x$ de la suite d'observations et la classe représentée par l'état e_3 contient le segment $y-y$. La segmentation est fondée sur la maximisation de la probabilité *a posteriori*.

classes sont respectivement décrites par les distributions $b_2(.)$ et $b_3(.)$ de la figure 2.16 page 32 si la segmentation est réalisée avec le HMM2 appris à l'issue de la première itération.

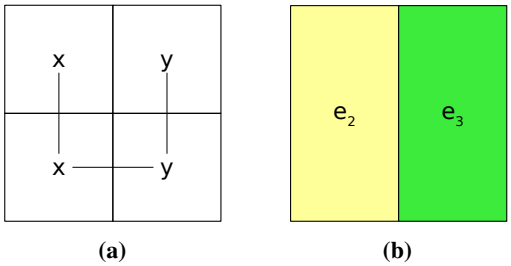


FIGURE 2.19 – Exemple de segmentation spatiale de 4 pixels voisins parcourus avec une courbe de type Hilbert-Peano qui les transforme en une suite monodimensionnelle d'observations : x-x-y-y.

2.3 Présentation et mise en œuvre d'ARPENTAGE

ARPENTAGE [3](Analyse de Régularités dans les Paysages : Environnement, Territoires et Agronomie) est un logiciel conçu pour la segmentation de territoires agricoles sur la base de l'évolution de leurs OCS. ARPENTAGE utilise et étend CARROTAGE : boîte à outils de fouille de données fondée sur les HMM2 et conçue pour la segmentation de séquences d'observations, que ces observations soient des OCS évoluant au fil des années (MARI et al., 1998 ; MARI et al., 1999) ou des nucléotides le long d'un ADN

3. http ://www.loria.fr/∼jfmari/App/.

bactérien (ENG et al., 2009). ARPENTAGE vient d'un besoin d'agronomes d'automatiser la spatialisation des résultats de la fouille temporelle des successions de cultures (MIGNOLET et al., 2007). Dans le cadre d'un travail de modélisation des systèmes de culture dans le bassin de la Seine en lien avec des problèmes de pollution diffuse des eaux souterraines et de surface de ce bassin, MIGNOLET et al. (2007) ont réalisé, à l'aide de CARROTAGE, des fouilles de données temporelles de triplets de cultures [4] sur chacune des 147 Petites Régions Agricoles (PRA) du bassin de la Seine. Ces auteurs ont ensuite effectué une analyse en composantes principales sur les 64 triplets de cultures majoritaires afin de réduire ces variables aux 20 premiers axes factoriels. Une classification ascendante hiérarchique effectuée sur le tableau des PRA et des 20 axes factoriels retenus leur a permis de distinguer 20 types de PRA dont la spatialisation a défini des zones homogènes correspondant à des groupes de PRA contiguës caractérisées par des combinaisons similaires de triplets de cultures (figure 2.20). La lourdeur de cette procédure et sa dépendance d'un maillage spatial prédéfini (les PRA) a motivé le développement d'ARPENTAGE en vue d'automatiser la spatialisation des régularités temporelles indépendamment d'un maillage prédéfini. ARPENTAGE a été développé pendant et pour cette thèse qui a permis de le tester et de le valider comme boîte à outils servant à la modélisation de territoires agricoles de dimension régionale sans recours à un maillage spatial prédéfini.

ARPENTAGE est un logiciel libre, écrit en C++, et fonctionnant en ligne de commande sous les systèmes d'exploitations Unix. Il est constitué d'un ensemble de programmes permettant :

1. de spécifier un modèle initial,
2. d'estimer ses paramètres,
3. de visualiser les transitions entre états ainsi que le contenu des états des HMM2, et
4. de fournir des shapefiles représentant la segmentation du territoire agricole selon la nature de l'observation et le nombre d'états définis par l'utilisateur.

Les résultats d'ARPENTAGE sont interprétés et validés par des spécialistes du domaine (agronomes, écologues, etc.).

Dans la suite de cette section nous décrivons les quatre principales étapes de fouille de données nécessaires pour réaliser des segmentations avec ARPENTAGE :

– la première étape consiste à préparer le corpus de données au format accepté par ARPENTAGE et à définir les observations dont la densité de probabilité constituera le contenu des états cachés du HMM2, ensuite vient :
– la spécification du modèle qui consiste à décrire la topologie du HMM2 ainsi que le contenu des états,
– le HMM2 initial est alors appris sur le corpus de la séquence d'observations, enfin
– les résultats de la fouille peuvent être visualisés à travers l'affichage des paramètres du modèle et des probabilités a posteriori des transitions et l'édition de cartes de zones homogènes du point de vue des densités d'observations associées aux états.

4. ***Triplets de cultures*** : successions de cultures de trois années.

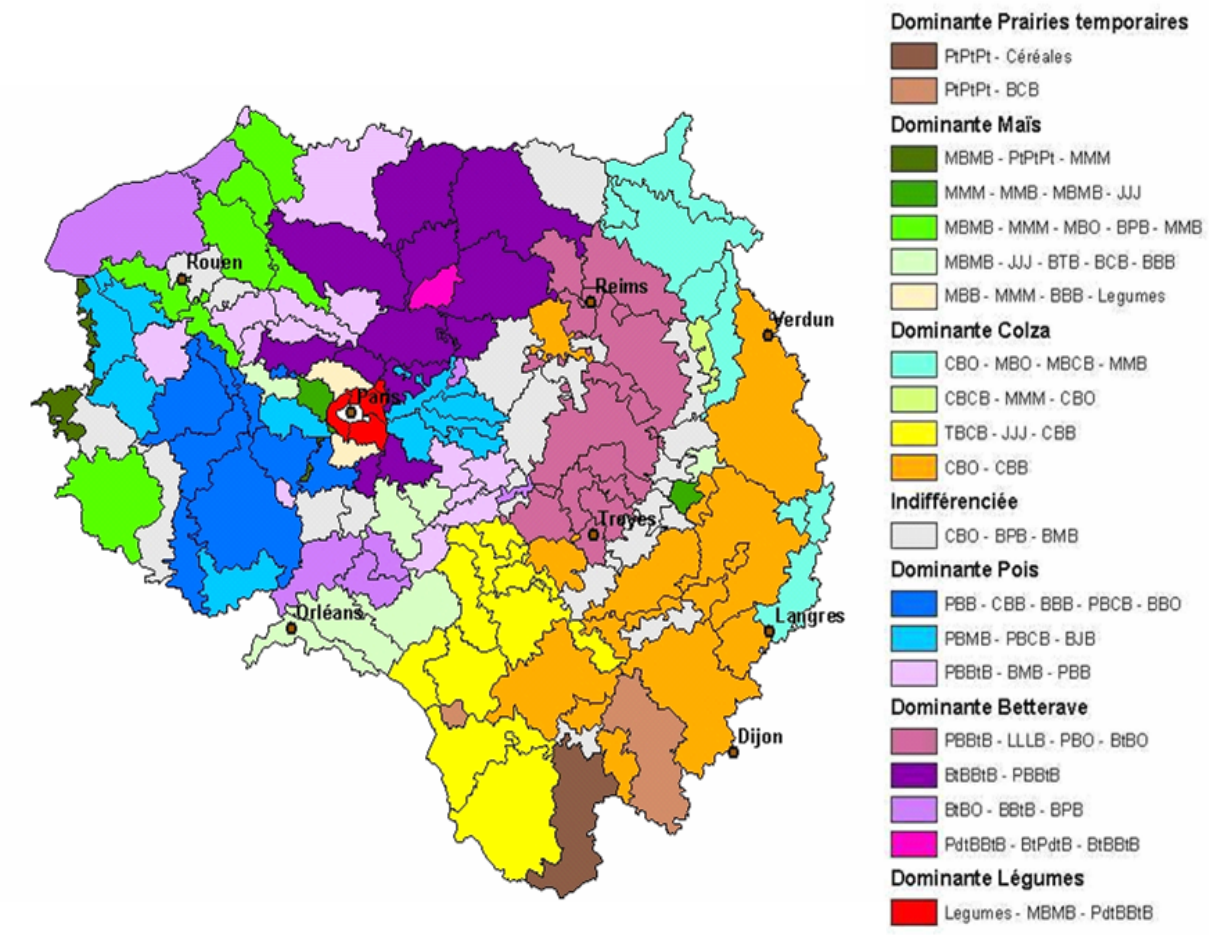


FIGURE 2.20 – Typologie des PRA du bassin de la Seine selon les successions culturales majoritaires pratiquées entre 1992 et 1998 (Source : MIGNOLET, 2008). Pt : prairies temporaires, B : blé, C : colza, M : maïs, P : pois, L : luzerne, Bt : betterave, Pdt : pomme de terre.

2.3.1 Préparations préalables des données

Préparation du corpus

Une préparation préalable du corpus de données est généralement nécessaire. ARPENTAGE prend en entrée une matrice de données discrètes (les codes d'OCS) où les lignes représentent les sites spatiaux échantillonnés suivant un maillage de points régulièrement espacés, et les colonnes représentent leurs valeurs temporelles (les OCS annuelles). Cette matrice de données constitue le corpus de la suite d'observations et se présente sous forme d'un fichier texte. Pour les segmentations où la composante spatiale entre en considération, ARPENTAGE réordonne la suite d'observations du corpus selon le parcours de Peano (*cf.* figure 2.10 de la section 2.2.5).

Définition des observations

Choix du type d'observation Le choix du type d'observation dépend de l'objectif de la fouille de données. ARPENTAGE permet de définir des types d'observations simples ou composites. Ces dernières sont constituées d'un ensemble d'OCS associés ou non à d'autres données catégorielles telles que des facteurs biophysiques comme le type de sol et la classe de distance par rapport à une entité paysagère, ou des faits de biodiversité comme l'occupation de la parcelle par un nid. Pour modéliser l'OTAA, nous traitons des observations constituées exclusivement d'OCS. L'observation choisie constitue la modalité d'un pixel de l'image représentant la mosaïque agricole. Plusieurs observations sont envisageables. Nous présentons ci-après les principales observations utilisées dans ARPENTAGE.

1. L'OCS en un point d'une parcelle représentant son occupation. Cette observation est pratique pour calculer l'assolement moyen et l'évolution temporelle de celui-ci.
2. La succession d'OCS en un point d'une parcelle sur deux ou plusieurs années successives se chevauchant temporellement. Cette observation est utile pour retrouver les successions dominantes selon la méthode de fouille développée par LE BER et al. (2006) et MIGNOLET et al. (2007).
3. Le quintuplé d'OCS défini par les OCS associées à 5 sites voisins dans la grille des points équidistants de l'échantillonnage : un site central et ses 4 voisins situés au Nord, au Sud, à l'Est et à l'Ouest. L'ensemble des quintuplés OCS est obtenu en parcourant point par point la grille d'échantillonnage en chevauchant les quintuplés spatialement. Les quintuplés permettent de calculer l'information mutuelle spécifique (PMI pour Pointwise Mutual Information) entre les OCS voisines deux à deux (*cf.* encadré 1), de tester l'isotropie du territoire d'étude (*cf.* encadré 2) et d'analyser le système de voisinage des parcelles en termes d'OCS lorsque le milieu est anisotrope. Nous choisirons les quintuplés dans l'étude des voisinages des OCS et des successions d'OCS (*cf.* section 3.3.2 page 65) pour vérifier l'isotropie du territoire d'étude de Chizé vis à vis des OCS.
4. Le couple (OCS, OCS d'une parcelle voisine) se chevauchant spatialement et le couple (succession d'OCS, successions d'OCS d'une parcelle voisine) se chevauchant temporellement et spatialement.

ARPENTAGE utilise la PMI(x,y) comme mesure du degré d'attraction ou de répulsion entre l'OCS x sur un site et l'OCS y sur le site voisin dans la courbe de Peano réduite ou entre les points voisins deux à deux des quintuplés d'OCS. La PMI est une mesure symétrique définie comme suit :

$$\text{PMI(x,y)} = \text{PMI(y,x)} = log\,\frac{P(y|x)}{P(y)} = log\,\frac{P(x|y)}{P(x)} = log\,\frac{P(x,y)}{P(x)\times P(y)}$$

La PMI(x,y) compare la probabilité *a priori* $P(y)$ de l'OCS y et la probabilité *a posteriori* $P(y|x)$. Symétriquement, la PMI(x,y) compare la probabilité *a priori* $P(x)$ de l'OCS x avec la probabilité *a posteriori* $P(x|y)$. Le signe de la PMI renseigne sur le résultats de cette comparaison :

- PMI(x,y)>0 signifie que la localisation en un point de l'OCS x ou y augmente la probabilité de localisation au voisinage de l'autre OCS. On dit qu'il y a une attraction entre x et y.
- PMI(x,y)<0 signifie que la localisation en un point de l'OCS x ou y diminue la probabilité de localisation au voisinage de l'autre OCS. On dit qu'il y a une répulsion entre x et y.
- PMI(x,y)$=0$ signifie que l'allocation de l'OCS x est indépendante de l'allocation de l'OCS y.

Encadré 1 – Définition de la PMI.

Le test d'isotropie sous ARPENTAGE se fait à l'aide d'un HMM2 linéaire utilisant des quintuplés (site central et ses 4 voisins situés au Nord, au Sud, à l'Est et à l'Ouest *cf.* section 2 page 37) d'OCS comme observations. A la dernière itération de l'algorithme Forward-Backward, les comptes de quintuplés sont calculés sur chaque état et permettent le calcul des comptes des cliques Nord (S,V_{No}), Sud (S,V_{So}), Est (S,V_{Es}) et Ouest (S,V_{We}), où S et V sont des variables aléatoires d'OCS, respectivement dans un site (S) et dans son site voisin (V). A partir de ces comptes, on estime les lois marginales $P(S,V)$ dans chaque direction.

La mosaïque agricole est isotrope si les distributions jointes $P(S,V)$ sont les mêmes quelle que soit la direction du voisinage.

La distance entre deux distributions est calculée à l'aide de la divergence :

$$div(f,g) = \frac{1}{2}\sum_x (f(x) - g(x))\,log\,\frac{f(x)}{g(x)} \tag{2.11}$$

où f et g sont deux distributions discrètes sur le même espace décrit par x.

Encadré 2 – Test d'isotropie avec ARPENTAGE.

Ces couples d'observations permettent, respectivement, de fouiller les voisinages entre OCS et successions d'OCS et leurs évolutions d'une façon efficace lorsque le milieu est isotrope vis-à-vis, respectivement, des OCS et des successions d'OCS (*cf.* encadré 3). Ces couples d'observations permettent aussi de calculer la PMI entre 2 OCS voisines.

Définition des observations dans ARPENTAGE L'utilisateur configure la nature de l'observation dans un fichier texte (dit fichier de configuration). Ce fichier est constitué de deux parties :

1. une table de correspondance qui fait le lien entre :
   - les codes d'OCS (tels qu'ils apparaissent dans le corpus de données),
   - le nom en clair correspondant à chaque code d'OCS (Forêt, Eau, Prairies, etc.),
   - et le numéro d'un éventuel regroupement attribué à chaque OCS ;
2. et une liste des noms des nouvelles catégories d'OCS (suite au regroupement).

Pour définir des observations constituées de plusieurs OCS (successions d'OCS, cliques temporelles, quintuplés d'OCS), il suffit de reproduire cette description dans le fichier de configuration autant de fois qu'il y a d'OCS dans l'observation. La figure 2.21 donne un exemple de fichier de configuration avec des observations consistant en des successions d'OCS sur deux années. ARPENTAGE possède un programme (`ter2indice`) qui construit à partir du corpus l'espace des possibles et génère une liste d'indices correspondant aux observations présentes dans le corpus en tenant compte de la définition donnée dans le fichier de configuration. Une ligne de commande typique permettant de générer cette liste d'observations est :

```
ter2indice corpus.txt -t fichierDeConfiguration.cfg -o listeDesObservations.lst
```

corpus.txt est le corpus d'OCS (*cf.* paragraphe 2.3.1) ;

fichierDeConfiguration.cfg est spécifié lors de la définition des observations ;

listeDesObservations.lst est le résultat de `ter2indice`

2.3.2 Spécification du HMM2 initial

ARPENTAGE peut utiliser des modèles de n'importe quelle topologie. Cependant, pour la modélisation de l'OTAA nous utilisons principalement deux types de modèles : les ***modèles linéaires*** et les ***modèles ergodiques***.

Un modèle linéaire est un HMM2 orienté gauche droite où les transitions possibles se font d'un état vers lui-même ou vers l'état suivant (figure 2.22 (a)). Les états du modèle linéaire initial sont généralement définis par des distributions d'observations équiprobables. Ce type de modèles permet de réaliser des segmentations temporelles.

Dans un modèle ergodique toutes les transitions entre états sont possibles (figure 2.22 (b)). Chaque état du modèle ergodique peut être associé :

Pour la prise en compte du système de voisinage irrégulier de la mosaïque agricole, ARPENTAGE implémente la méthode décrite en section 2.2.5 (page 26, particularité 3) qui consiste à utiliser le parcours d'une courbe de Peano réduite en ajustant la profondeur de la fractale à la taille des parcelles élémentaires.

Clique

Deux sites voisins sur la fractale réduite définissent une *clique* si ces sites sont associés à des observations différentes. Cela revient à n'échantillonner le territoire que le long des frontières des parcelles élémentaires occupées par des observations (OCS ou successions d'OCS) différentes. La fréquence des cliques est proportionnelle à la longueur des frontières. Cette proportionnalité est d'autant plus précise que la résolution d'échantillonnage est plus fine. La fouille du voisinage consiste à estimer les fréquences de ces couples d'observations :
– en parcourant la courbe de Peano réduite clique par clique, et pour chaque clique,
– en parcourant la période d'étude année par année.

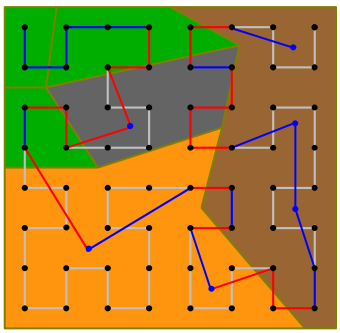
Clique temporelle

Pour souligner la prise en compte de la dimension temporelle dans l'estimation des fréquences du couple (observation, observation voisine), celui-ci sera appelé *clique temporelle*. Plus particulièrement,
– le couple (OCS, OCS d'une parcelle voisine) sera appelé *clique temporelle d'OCS*,
– et le couple (succession d'OCS, successions d'OCS d'une parcelle voisine) sera appelé *clique temporelle de successions d'OCS*.

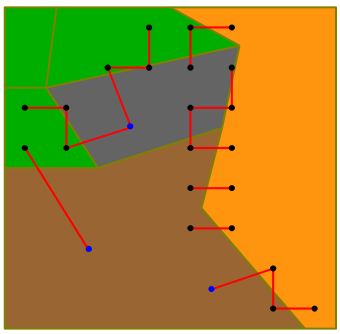
Notons toutefois que ces appellations comportent un certain abus de langage dans la mesure où une clique est bien identifiée spatialement, alors que le couple (observation, observation voisine) n'est pas spécifique à une clique, il s'agit plutôt d'une estimation de la fréquence des cliques ayant la même configuration dans l'ensemble du territoire d'étude.

Les figures (a) et (b), ci-contre, illustrent la méthode de définition des cliques moyennant l'ajustement, à la surface des parcelles, de la profondeur de la courbe fractale de Peano. Chaque motif de la courbe fractale de Peano se trouvant à l'intérieur d'une parcelle est réduit à son centre de gravité. Par exemple, dans la parcelle du bas (figure (a)), un motif de 16 points et un autre de 4 points sont réduits. Les cliques « plein champ » et les cliques entre deux parcelles associées à la même observation sont ignorées. Les cliques retenues sont les segments qui se trouvent le long des frontières entre deux parcelles.
Le tableau ci-dessous donne les fréquences des cliques temporelles d'OCS calculées à partir des deux assolements successifs (figures (a) et (b)). Chaque couleur correspond à une OCS.

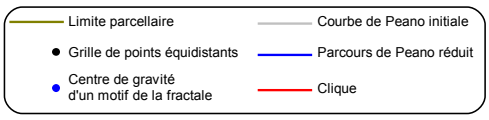
Cliques temporelles d'OCS	Fréquences
(vert, gris)	$14/36$
(vert, orange)	$2/36$
(orange, marron)	$12/36$
(orange, gris)	$3/36$
(gris, marron)	$3/36$
(vert, marron)	$2/36$



(a) Assolement d'une année (t)



(b) Assolement de l'année suivante ($t+1$)



Encadré 3 – Méthode de prise en compte des voisinages des OCS et des successions d'OCS dans ARPENTAGE.

```
{
{"Forêt ", 1, 0},
{"Eau", 2, 0},
{"Bâti", 3, 0},
{"Céréales", 4, 1},
{"Maïs", 5, 2},
{"Prairies", 6, 3},
}
```
(a)

```
{
{"NonAgricole"},
{"Céréales"},
{"Maïs"},
{"Prairies"}
}
```
(b)

```
{
{"Forêt ", 1, 0},
{"Eau", 2, 0},
{"Bâti", 3, 0},
{"Céréales", 4, 1},
{"Maïs", 5, 2},
{"Prairies", 6, 3},
}
{
{"NonAgricole"},
{"Céréales"},
{"Maïs"},
{"Prairies"}
}
```
(c)

FIGURE 2.21 – Exemple d'un fichier de configuration tiré de l'étude du bassin versant du Yar et spécifiant des observations sous forme de successions biennales d'ocs. Ce fichier est constitué de trois structures illustrées dans (a) et (b). (a) table de correspondance qui fait le lien entre : (i) les codes d'ocs tels qu'ils apparaissent dans le corpus de données (dans cet exemple : 1 à 6), (ii) le type d'ocs correspondant à chaque code (Forêt, Eau, etc.) et (iii) le numéro d'un éventuel regroupement attribué à chaque ocs. Après regroupement, les catégories d'ocs sont : NonAgricole, Céréales, Maïs et Prairies. (b) liste des noms des nouvelles catégories d'ocs (suite au regroupement). (c) Les ensembles d'occupations apparaissent deux fois car l'observation est constituée de deux ocs.

– à une distribution nulle sauf pour un nombre restreint d'observations, dans ce cas l'état est dit de ***Dirac***
– ou à une distribution équiprobable des observations et dans ce cas l'état est dit de ***réserve***.

Des combinaisons entre ces topologies sont utilisées pour construire des HMM2 hiérarchiques où chaque état du modèle ergodique est un HMM2 linéaire ou ergodique et permettent de réaliser des segmentations à la fois temporelles et spatiales.

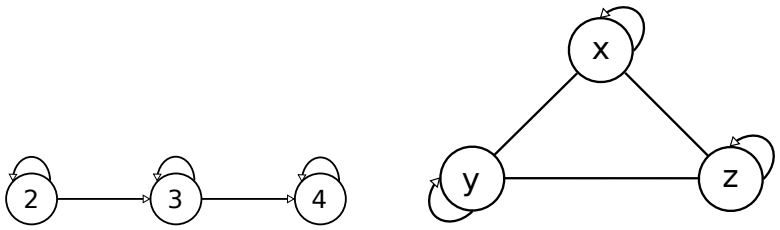


FIGURE 2.22 – Topologies de base d'HMM2 utilisés pour la modélisation de l'OTAA avec ARPENTAGE. (a) Exemple d'un HMM2 linéaire à 3 états effectifs. Cette topologie est utilisée pour la segmentation temporelle. (b) Exemple d'un HMM2 ergodique à 3 états effectifs. Cette topologie est utilisée pour la segmentation spatiale ou pour afficher les transitions entre les observations dans un diagramme de Markov. Les deux états de début et l'état final ne sont par représentés.

ARPENTAGE possède un programme (`editmodel`) qui crée le modèle initial (modèle.mod) à partir du descriptif (descriptifDuModèle.lst) et du fichier de configuration fourni en argument comme dans la ligne de commande suivante :

```
editmodel -t fichierDeConfiguration.cfg -i listeDesObservations.lst -d descriptifDuModèle.lst -o modèle.mod
```

fichierDeConfiguration. cfg est spécifié lors de la définition des observations (*cf.* section 2.3.1),

listeDesObservations. lst est le résultat de l'exécution du programme `ter2indice` (cf. section 2.3.1),

descriptifDuModèle. lst est un fichier texte décrivant les transitions entre états (topologie) et le contenu des états du HMM2 initial. L'utilisateur décrit le modèle initial dans ce fichier composé de deux parties. La première partie précise la topologie du modèle au moyen d'une liste de trois entiers séparés par un espace. Les transitions entre des états est définie suivant la structure suivante : « numéro de l'état de départ » « numéro de l'état d'arrivée » « poids donné à la transition ». La deuxième partie définit les distributions des observations initiales associées à chaque état. Des exemples de fichiers de description pour différents HMM2 sont donnés dans la figure 2.23.

2.3.3 Apprentissage du modèle initial

Les paramètres du modèle initial sont alors estimés de manière itérative sur le corpus de la suite d'observations en utilisant un programme (`fwXinra`) qui implémente l'algorithme de Forward-Backward. Une ligne de commande typique est :

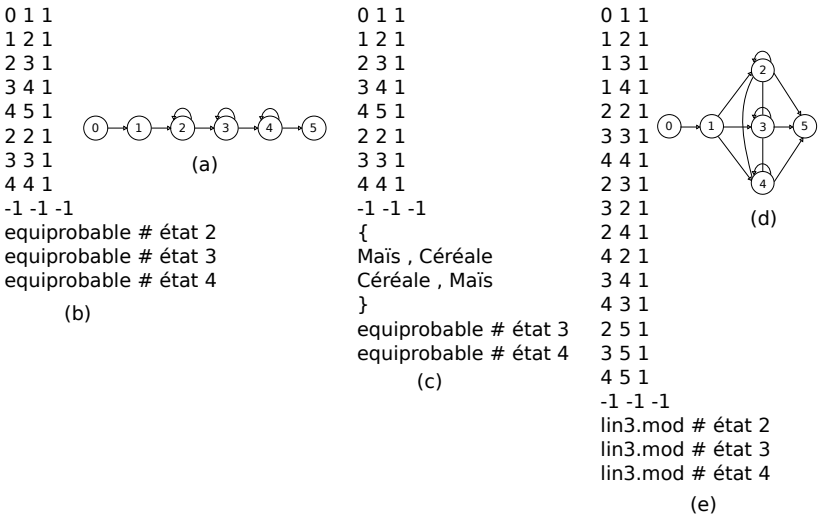


FIGURE 2.23 – Exemples de fichiers de description de différents HMM2. Les premières lignes décrivent les transitions (une ligne est structurée : « Etat-Départ » « Etat-Arrivée » « Poids »), les dernières lignes décrivent les distributions associées aux états. La séparation entre les deux parties est marquée par la ligne : $-1\ -1\ -1$. Seules les distributions associées aux états effectifs (nommés 2, 3 et 4) sont spécifiées. Les états initiaux (0 et 1) et l'état final ne sont pas décrits dans la deuxième partie. (a) Topologie d'un HMM2 linéaire à 3 états effectifs (noté `lin3`). (b) fichier de description d'un `lin3` dont les états sont des distributions uniformes d'observations. (c) Fichier de description d'un `lin3` où l'état 2 est destiné à capturer les successions biennales entre Maïs et Céréale. Ces deux observations ont une probabilité de 0,5 alors que les probabilités de tous les autres couples sont nulles. (d) Topologie d'un HMM2 ergodique à 3 états effectifs (noté `ergo3`). (e) fichier de description d'un `ergo3` dont les états sont des `lin3`. Dans la deuxième partie du fichier de description, `lin3.mod` est un fichier dans le répertoire courant issu du résultat de l'exécution du programme `editmodel`.

```
fwXinra -n NombreDItérations modèle.mod corpus.txt -t fichierDeConfiguration.cfg -o modèle.mod1
```

Le modèle initial (modèle.mod) est estimé en utilisant le corpus des observations contenu dans le fichier corpus.txt.

La détermination du nombre d'itérations est un problème statistique qui ne possède pas de solution exacte. Pour aider l'utilisateur à le définir, ARPENTAGE calcule, à chaque itération, le gain de vraisemblance par la différence des moyennes géométriques de la vraisemblance [5] avant et après l'itération. Le nombre d'itérations est un paramètre fixé par l'utilisateur qui fait un compromis entre le temps nécessaire pour chaque itération et le gain de vraisemblance après chaque itération. Le modèle appris qui en résulte est stocké dans le fichier de sortie (modèle.mod1).

2.3.4 Visualisation des résultats

ARPENTAGE possède des programmes permettant la visualisation des résultats de la segmentation. La segmentation, réalisée avec l'algorithme CEM, est opérée à la dernière itération. Pour chaque observation o_t, ARPENTAGE calcule différents types de probabilités *a posteriori*, dont 2 nous intéressent particulièrement :

- la probabilité *a posteriori* de rester dans l'état e_i à l'indice t sachant toute la suite d'observations o_1^T. Ce type permet de réaliser une segmentation spatiale.
- la probabilité *a posteriori* de la transition $e_i \rightarrow e_j$ à l'indice t sachant la suite des observations o_1^T. Ce type permet de visualiser les transitions entre les observations (OCS ou successions d'OCS) dans un diagramme de Markov.

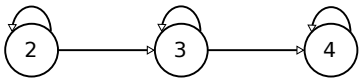
Dans notre cas, la suite d'observations est la séquence représentée par l'ensemble des points de la grille d'échantillonnage. Dans le cas de la fouille spatiale, ces points sont ordonnés suivant le parcours de la courbe fractale de Peano.

ARPENTAGE permet d'afficher les distributions du modèle (figure 2.24), les probabilités *a posteriori* de transitions entre les états (figure 2.25) et de générer des shapefiles permettant de visualiser les résultats des segmentations spatiales (figure 2.26).

2.4 Conclusion

Nous avons présenté dans ce chapitre les deux territoires agricoles étudiés ainsi que leurs bases de données d'OCS. Ce sont deux territoires agricoles aux propriétés contrastées. Le site d'étude de Chizé est un territoire régional de dimension moyenne où les informations d'OCS sont issues de relevés de terrain quasi exhaustifs pendant la période d'étude. Le bassin versant du Yar est un petit territoire régional où les informations d'OCS sont d'origine satellitaires. Les corpus de données d'OCS issus de ces deux régions agricoles donnent la localisation et l'évolution des OCS sur une décennie.

5. $\sqrt[T]{L(o)}$ est la moyenne géométrique de la vraisemblance pour une séquence de longueur T, *cf.* équation 2.1.



Prairies	0,31	Prairies	0,29	Blé	0,29
Blé	0,22	Blé	0,26	Prairies	0,27
Orge	0,16	Colza	0,14	Colza	0,17
Colza	0,12	Orge	0,11	Orge	0,12
Maïs	0,07	Maïs	0,08	Maïs	0,06
Jachères	0,05	Jachères	0,05	Vergers	0,02

FIGURE 2.24 – Exemple de visualisation des distributions associées aux états d'un HMM2 (LE BER et al., 2006). Le HMM2 utilisé dans la fouille est de type linéaire à 3 états. Lors de la spécification du modèle, les états étaient initialisés avec des OCS équiprobables. Après l'apprentissage sur les données issues d'une région agricole, les distributions de cultures obtenues mettent en valeur une tendance de régression des Prairies au profit du Blé.

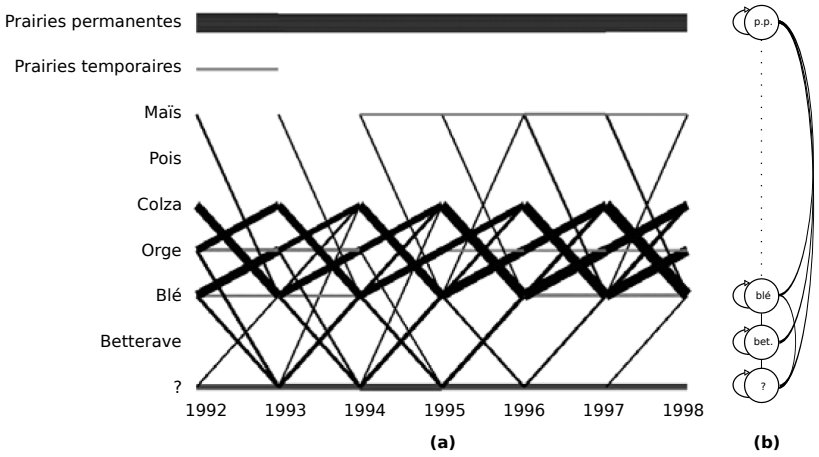


FIGURE 2.25 – Exemple d'un diagramme de Markov affiché par ARPENTAGE après une fouille temporelle de données d'OCS du Plateau Barrois (en Lorraine) issu de l'étude de MIGNOLET et al. (2004). (a) L'utilisateur peut visualiser les probabilités *a posteriori* des transitions entre les états (lignes diagonales et horizontales) dont l'épaisseur est proportionnelle à la probabilité. (b) Le HMM2 utilisé est de type ergodique à 9 états dont 8 états sont des états de Dirac, caractérisés par une seule observation par état, et un état dit de réserve, marqué par « ? ». Dans cet état de réserve, les observations étaient équiprobables avant l'apprentissage. Après l'apprentissage, cet état capture les OCS non définies dans les états de Dirac.

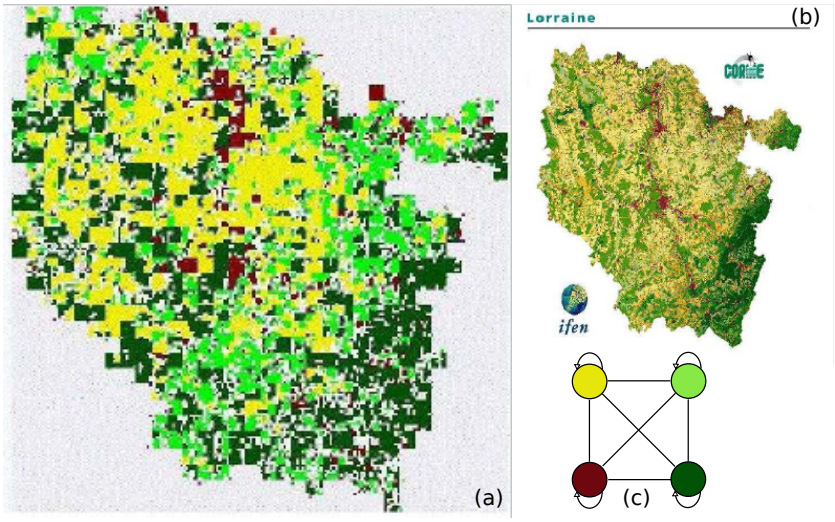


FIGURE 2.26 – Exemple de visualisation d'une segmentation spatiale (MARI et LE BER, 2006). (a) segmentation en quatre classes spatiales des données *Teruti* d'une région agricole en Lorraine. (b) Image satellite de la même région avec une résolution quatre fois plus élevée. (c) Topologie du HMM2 utilisé dans cette segmentation. Chaque état du HMM2 représente une classe spatiale.

Pour étudier la mosaïque agricole dans ses composantes spatiale et temporelle, nous nous sommes intéressés aux approches Markoviennes. Nous avons utilisé une technique de fouille de données avec des modèles de Markov cachés initialement mise au point pour la segmentation d'images numériques. Cette technique approxime un champ de Markov par une chaîne de Markov en transformant les deux dimensions du plan en une séquence monodimensionnelle par un parcours de type fractal du plan. Cette approximation se justifie par le fait que cette méthode a montré de bons résultats comparée aux méthodes mettant en œuvre des champs 2D d'une façon plus rigoureuse. Un autre avantage pratique de cette approximation est que les algorithmes d'estimation et de classification sont les mêmes pour traiter les composantes temporelle et spatiale. Au fil des travaux de notre équipe de recherche, cette technique a été enrichie pour mieux l'adapter aux particularités du contexte agricole. Une adaptation majeure testée et validée dans notre travail consiste à automatiser la spatialisation des régularités temporelles indépendamment d'un maillage prédéfini. ARPENTAGE a été développé pendant et pour cette thèse et celle-ci l'a testé et l'a validé comme boîte à outils servant à la modélisation de territoires agricoles de dimension régionale sans recours à un maillage spatial prédéfini.

Le chapitre suivant présente une synthèse de notre méthode de modélisation de l'OTAA avec ARPENTAGE que nous avons testée et validée sur les deux territoires d'étude.

2.5 Références

ABRIAL, D, L AZIZI, M CHARRAS-GARRIDO et F FORBES (2010). « Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie à l'aide de champs de Markov cachés ». Dans : *42èmes Journées de Statistique*. Marseille, France, France. URL : http://hal.inria.fr/inria-00494838.

AZIZI, L, F FORBES, S DOYLE, M CHARRAS-GARRIDO et D ABRIAL (2011). *Spatial risk mapping for rare disease with hidden Markov fields and variational EM*. Anglais. Rapport de recherche RR-7572. INRIA.

BAUM, LE, T PETRIE, G SOULES et N WEISS (1970). « A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains ». Dans : *The annals of mathematical statistics* 41.1, p. 164–171.

BENMILOUD, B et W PIECZYNSKI (1995). « Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images ». Dans : *Traitement du signal* 12.5, p. 433–454.

BERTHOD, M, Z KATO, S YU et J ZERUBIA (1996). « Bayesian image classification using Markov random fields ». Dans : *Image and Vision Computing* 14.4, p. 285–295.

CASTELLAZZI, MS, GA WOOD, PJ BURGESS, J MORRIS, KF CONRAD et JN PERRY (2008). « A systematic representation of crop rotations ». Dans : *Agricultural Systems* 97.1-2, p. 26–33.

CELEUX, G et G GOVAERT (1992). « A classification EM algorithm for clustering and two stochastic versions ». Dans : *Computational Statistics & Data Analysis* 14.3, p. 315–332.

CORGNE, S (2004). « Hiérarchisation des facteurs de changements de l'occupation hivernale des sols: Application au bassin versant du Yar (Bretagne) ». Dans : *Norois* 193, p. 17–29.

ENG, C, C ASTHANA, B AIGLE, S HERGALANT, J-F MARI et P LEBLOND (2009). « A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods ». Dans : *Journal of Computational Biology* 16.9, p. 1211–1225.

FINE, S, Y SINGER et N TISHBY (1998). « The Hierarchical Hidden Markov Model: Analysis and Applications ». Dans : *Machine Learning* 32, p. 41–62.

JELINEK, F (1976). « Continuous speech recognition by statistical methods ». Dans : *Proceedings of the IEEE* 64.4, p. 532–556.

KROGH, A (1998). « Chapter 4 An introduction to hidden Markov models for biological sequences ». Dans : *Computational Methods in Molecular Biology*. Sous la dir. de SL SALZBERG, DB SEARLS et S KASIF. T. 32. New Comprehensive Biochemistry. Elsevier, p. 45 –63.

LE BER, F, M BENOÎT, C SCHOTT, J-F MARI et C MIGNOLET (2006). « Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software ». Dans : *Ecological Modelling* 191.1, p. 170–185.

MARI, J-F, J-P HATON et A KRIOUILE (1997). « Automatic Word Recognition Based on Second-Order Hidden Markov Models ». Dans : *IEEE Transactions on Speech and Audio Processing* 5, p. 22–25.

MARI, J-F et F LE BER (2006). « Temporal and Spatial Data Mining with Second-Order Hidden Markov Models ». Dans : *Soft Computing*. ISSN:1432-7643 10.5, p. 406–414.

MARI, J-F, F LE BER et M BENOÎT (1998). « Reconnaissance de successions culturales par modèles de Markov : une étude préliminaire ». Dans : *Journées Cassini*. Marne-la-Vallée.

— (1999). « Classification de successions culturales par modèles de Markov ». Dans : *Septième journées de la Société Francophone de Classification - SFC'99*. Colloque avec actes et comité de lecture, p. 177–184.

MIGNOLET, C (2008). « Modélisation de l'organisation spatiale des systèmes agricoles et de son évolution dans des démarches d'appui au développement ». Thèse de doct. Institut des sciences et industries du vivant et de l'environnement (Agro Paris Tech).

MIGNOLET, C, C SCHOTT et M BENOÎT (2004). « Spatial dynamics of agricultural practices on a basin territory: a retrospective study to implement models simulating nitrate flow. The case of the Seine basin ». Dans : *Agronomie* 24.4, p. 219–236.

— (2007). « Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale ». Dans : *Science of the Total Environment* 375.1-3, p. 13–32.

RABINER, LR (1989). « A tutorial on hidden Markov models and selected applications in speech recognition ». Dans : *Proceedings of the IEEE* 77.2, p. 257–286.

WELCH, LR (2003). « Hidden markov models and the baum-welch algorithm ». Dans : *IEEE Information Theory Society Newsletter* 53.4, p. 1–10.

Chapitre 3

Notre méthode de modélisation de l'OTAA avec ARPENTAGE

Dans ce chapitre nous présentons une méthode pour modéliser et décrire l'OTAA à des échelles compatibles avec les échelles de déroulement des processus écologiques et environnementaux. Nous nous sommes appuyés sur le concept de succession de cultures pour modéliser les dynamiques de l'OTAA dans 2 territoires agricoles de dimensions régionales : le site d'étude de Chizé et le bassin versant du Yar. Ce chapitre reprend tous les éléments méthodologiques mobilisés dans nos articles. Notre méthode de modélisation part de l'hypothèse que l'OTAA est un champ de Markov de successions. Nous validons cette hypothèse à travers trois contributions principales :

1. la description de l'OTAA par segmentation temporo-spatiale [1] du territoire agricole en zones compactes (patches) caractérisées par des distributions de régularités temporelles d'OCS et de successions d'OCS ;
2. le développement d'une méthode de fouille des voisinages des successions d'OCS ;
3. l'articulation des régularités révélées par notre approche de fouille de données à l'échelle régionale avec des règles identifiées par des experts en agronomie et en écologie à des échelles plus locales en vue d'expliquer les régularités et de valider les hypothèses des experts dans leurs domaines.

Avant d'exposer les grandes lignes de ces contributions, nous présentons des statistiques communes à notre démarche de fouille de données. Celles-ci sont opérées sur le corpus afin de paramétrer les modèles de fouille. Il s'agit d'analyses statistiques qui nous ont permis de choisir la résolution de l'échantillonnage spatial et la longueur de la succession d'OCS.

1. L'approche ***temporo-spatiale*** sera présentée en section 3.2.

TABLEAU 3.1 – Composition et fréquences cumulées des catégories d'ocs définies dans une première fouille exploratoire pour le site d'étude de Chizé (LAZRAK et al., 2009).

Catégorie d'ocs	ocs	Fréquence cumulée
Blé (B)	Blé, blé barbu, céréale	0,337
Tournesol (T)	Tournesol, ray-grass suivi Tournesol	0,476
Colza (C)	Colza	0,600
Urbain (U)	Bâti, péri-village, route	0,696
Prairies et Luzernes (P)	Prairie permanente, prairie année 1, prairie temporaire (2-3 ans), prairie âge inconnu, luzerne 1 an, luzerne 2 ans, luzerne 3 ans, luzerne $>$ 3 ans	0,774
Maïs (M)	Maïs, ray-grass suivi maïs	0,850
Forêts et friches (F)	Forêt ou haie, friche	0,884
Orge d'hiver (O)	Orge d'hiver	0,918
Ray-grass (R)	Ray-grass, ray-grass suivi ray-grass	0,942
Pois (S)	Pois	0,964
Autres (A)	Orge de printemps, vigne, jachère spontanée juin, moha, lin, avoine, Trèfle, fèverole, ray-grass suivi labour, ray-grass suivi inconnu, jachère spontanée suivie labour, mélange céréale légumineuse, culture printemps, moutarde, jardin/culture maraîchère, Sorgho/Millet, Sorgho, Millet, Labour, Tabac, Autre culture	1,000

3.1 Statistiques sur le corpus de données

3.1.1 Définition de l'observation élémentaire

Lorsque les modalités des ocs sont nombreuses, il est utile de construire une typologie d'ocs afin d'en diminuer le nombre et de simplifier la lisibilité des résultats de fouille de données avec ARPENTAGE. Cette étape se justifiait pour le cas du site d'étude de Chizé où le nombre de modalités d'ocs était relativement important (*cf.* section 2.1.2). Dans une première fouille exploratoire du site d'étude de Chizé (LAZRAK et al., 2009), nous avons suivi une démarche de typologie tenant compte des fréquences des ocs et de la similitude des conduites culturales (tableau 3.1). À chaque exercice de fouille de données avec ARPENTAGE, la typologie définissant les observations élémentaires peut être entièrement ou partiellement revue suivant l'objectif de la modélisation. Par exemple, nous avons adapté cette typologie pour modéliser l'évolution du voisinage des prairies en individualisant celles-ci de la catégorie « Prairies et Luzernes » (LAZRAK et al., 2011). Ci-après, les catégories d'ocs seront parfois appelées ocs pour simplifier.

3.1.2 Choix de la longueur des successions

L'interprétabilité des résultats de fouille utilisant des successions d'ocs dépend de la longueur des successions. Cette longueur peut être définie en faisant un compromis qui tient compte des avantages et inconvénients des longues successions. En effet, plus les successions sont longues, plus celles-ci traduisent les règles logiques que les agriculteurs mettent en œuvre pour définir l'ordre des cultures qu'ils

allouent aux parcelles de leurs exploitations. Mais les longues successions compliquent l'interprétation des résultats de fouille en raison de la difficulté de lisibilité d'une longue succession et du nombre important des successions différentes.

Nous avons défini une méthode qui aide à choisir une longueur à la fois courte et pertinente de successions d'OCS. Cette méthode compare la diversité des successions d'OCS en données réelles et en données générées aléatoirement pour différentes longueurs de successions. La diversité des successions d'OCS est évaluée à l'aide du programme `ter2indice` d'ARPENTAGE qui calcule le nombre d'observations différentes dans un corpus de données pour un type d'observation défini dans le fichier de configuration (*cf.* section 2.3.1). Afin de focaliser sur les logiques exprimées par les agriculteurs à travers la définition de l'ordre des cultures dans les successions, les OCS non agricoles ne sont pas tirées au hasard, et les cultures sont tirées au hasard en respectant les proportions réelles des cultures. Ces proportions sont calculées à partir de la moyenne des surfaces cultivées dans l'ensemble du territoire durant la période d'étude.

Pour les deux territoires d'étude, nous pouvons voir dans la figure 3.1 que les successions de 4 ans (notées $t, \ldots, t+3$) commencent à se différencier de la situation aléatoire. Ceci a appuyé notre choix de la succession quadriennale comme observation pour modéliser l'organisation des successions d'OCS dans le site d'étude de Chizé (LAZRAK et al., 2009).

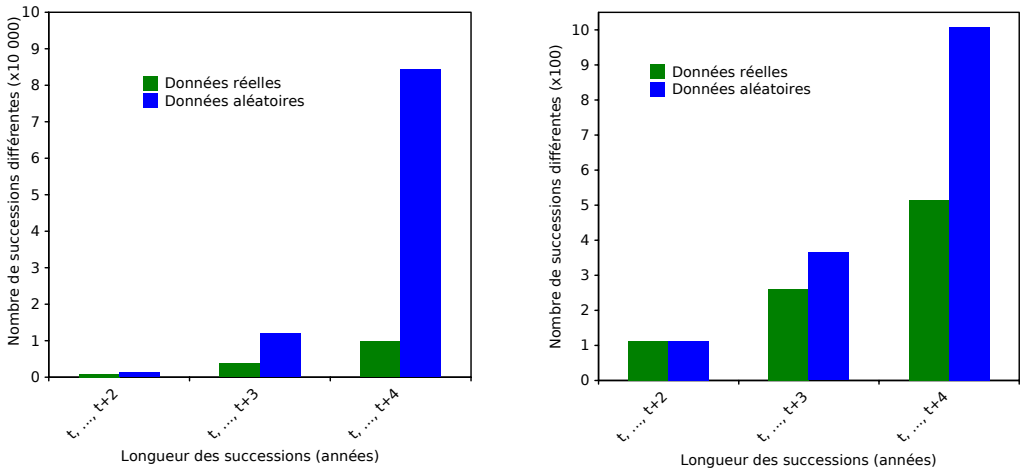


FIGURE 3.1 – Comparaison de la diversité des successions d'OCS entre les données réelles d'OCS et des données générées aléatoirement pour différentes longueurs de successions et pour les deux cas d'études : le site d'étude de Chizé (a) et le bassin versant du Yar (b). Le corpus des données aléatoires respecte les proportions réelles des cultures. Ces proportions sont calculées pour chaque cas d'étude à partir de la moyenne des surfaces cultivées durant la période d'étude.

3.1.3 Choix de la résolution d'échantillonnage

Pour les territoires agricoles de grandes étendues, une haute résolution d'échantillonnage génère de grands corpus de données qui augmentent les temps de calcul. Inversement, avec une résolution d'échantillonnage grossière, l'information relative aux petites parcelles est perdue. Afin de disposer d'un critère

objectif permettant de choisir une résolution spatiale appropriée, nous avons calculé le nombre des observations pour des résolutions d'échantillonnage de plus en plus grossières. La diversité des observations est évaluée à l'aide du programme `ter2indice` d'ARPENTAGE (*cf.* section 2.3.1 page 39).

La figure 3.2 illustre l'évolution de la diversité de deux types d'observations en fonction de la résolution de l'échantillonnage pour le site d'étude de Chizé. La résolution de 80 *m* permet d'obtenir un corpus 64 fois plus petit que le corpus initial avec une perte de seulement 6 % en termes de la diversité des successions, alors que pour les quintuplés (abordés et définis au paragraphe 2 page 37) d'OCS cette résolution en maximise la diversité. Entre 10 *m* et 80 *m*, les quintuplés atteignent plus de voisinages qu'ils n'en omettent, ce qui explique leur diversité croissante. Aux plus fines résolutions, les quintuplés épousent les frontières des parcelles et permettent d'interpréter les résultats proportionnellement à la longueur des frontières. Pour permettre cette proportionnalité, les quintuplés « plein champ » et les quintuplés situés entre des parcelles associées à la même observation sont ignorés. Les quintuplés retenus sont ceux qui se situent le long des frontières entres parcelles, d'où la proportionnalité avec la longueur des frontières pour les fines résolutions.

Le grand nombre de quintuplés générés ($>$ 20 000 à 10 $m \times 10$ m de résolution et $>$ 40 000 à 80 $m \times 80$ m) complique l'interprétation des résultats de la fouille . Pour fouiller les relations de voisinages, les quintuplés sont avantageusement remplacés par les cliques (*cf.* encadré 3 page 40) à condition de vérifier l'isotropie du territoire d'étude vis-à-vis de l'observation considérée. ARPENTAGE permet de vérifier l'isotropie du territoire suivant la démarche décrite dans l'encadré 2 page 38.

Le problème de choix de la résolution d'échantillonnage ne s'est pas posé pour le cas du bassin versant du Yar. La résolution initiale (20 $m \times 20$ m) a généré un corpus de données d'environ 150 000 points dont le traitement avec ARPENTAGE ne posait pas de contraintes de temps de traitement et de saturation de la mémoire comme pour le cas du corpus du site d'étude de Chizé.

3.2 Segmentation temporo-spatiale du territoire agricole

Notre méthode de segmentation ***temporo-spatiale*** du territoire agricole est une procédure de fouille de données réalisée en deux étapes. La première étape cherche à identifier des régularités temporelles sur l'ensemble du territoire sans tenir compte de leur localisation. La deuxième étape localise ces régularités en segmentant le territoire agricole en patches décrits par des distributions probabilistes des régularités temporelles. Il s'agit donc d'une démarche temporo-spatiale permettant de cartographier des objets temporels, contrairement à la démarche spatio-temporelle qui suit les trajectoires d'objets spatiaux au travers d'une série temporelle de cartes ou plus généralement d'images. Nous présentons dans cette section une synthèse de cette méthode de fouille de données que nous avons mis en œuvre sur les deux territoires d'études.

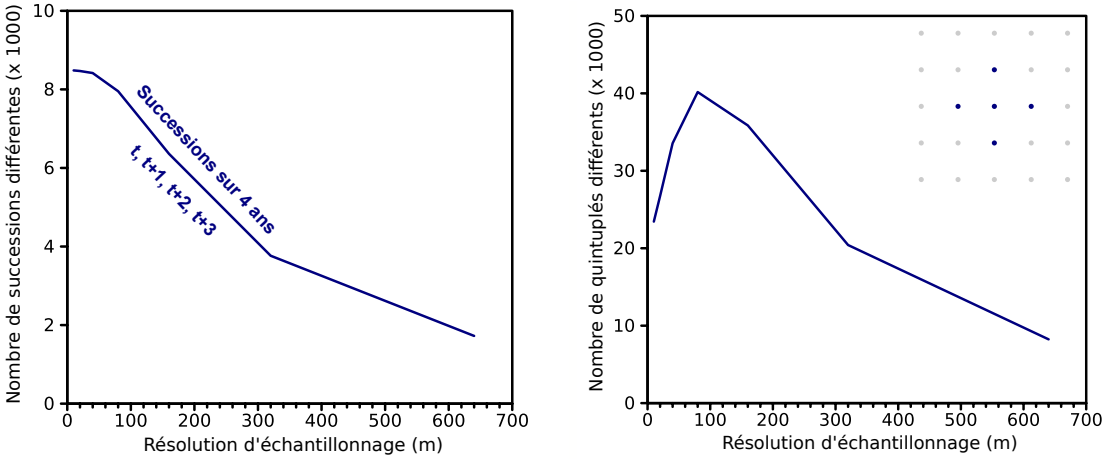


FIGURE 3.2 – Évolution de la diversité des observations en fonction de la résolution de l'échantillonnage. (a) Nombre de successions d'OCS (sur 4 ans) suivant la résolution spatiale d'échantillonnage. La courbe montre une perte progressive en diversité de successions d'OCS. (b) Nombre de quintuplés d'OCS suivant la résolution spatiale d'échantillonnage. La résolution spatiale $80m \times 80m$ donne la plus grande diversité de voisinages.

3.2.1 Recherche des régularités temporelles

La première étape de la segmentation temporo-spatiale est un exercice exploratoire de fouille de données qui vise à identifier des régularités [2] temporelles sans tenir compte de leur localisation. Pour ce faire, nous envisageons des fouilles temporelles complémentaires et de complexité croissante où chaque étape de fouille vise à rajouter « une brique de connaissance » aidant à la compréhension de l'organisation de l'activité agricole dans le territoire étudié.

Recherche des régularités d'évolution de l'assolement

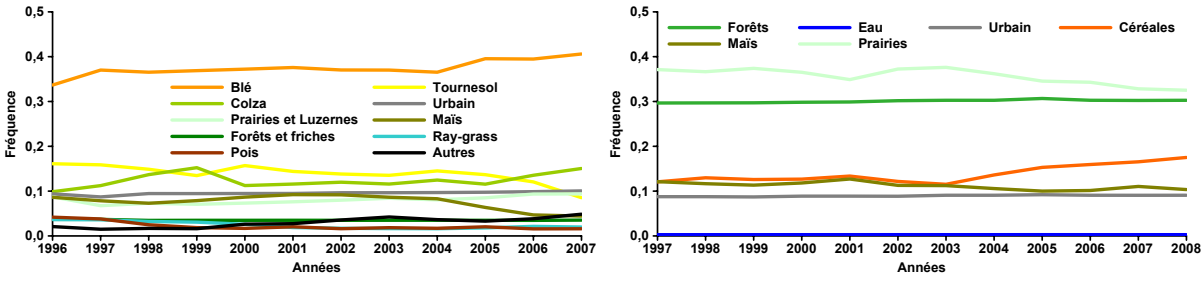
Ce type d'analyse temporelle a pour objectif de fournir une première vue d'ensemble sur l'OTAA à travers l'assolement global du territoire d'étude et son évolution durant la période d'étude. Cette analyse est fondée sur des observations élémentaires (OCS ou catégories d'OCS) pour lesquelles l'apprentissage du HMM2 permet d'estimer l'évolution des fréquences le long de la période d'étude. Les observations élémentaires sont choisies en fonction des critères présentés au début de ce chapitre (*cf.* section 3.1.1 page 50). La segmentation temporelle est réalisée avec un HMM2 linéaire (*cf.* section 2.3.2 page 39) dont le nombre d'états dépend du nombre de segments temporels recherchés. Les états sont initialisés avec des distributions uniformes des observations élémentaires (états équiprobables). Si le HMM2 est défini avec un seul état, le résultat de la fouille est l'assolement moyenné sur la période d'étude (*cf.* p.ex. tableau 3.1 page 50). Si le nombre d'états du HMM2 est égal au nombre d'années de la période d'étude, le résultat obtenu est l'évolution de l'assolement du territoire d'étude qui permet de repérer d'éventuelles dates de « ***ruptures*** » [3]

2. Le terme « ***régularité*** » a été défini en section 2.2.1 page 22.

3. Par « ***date (ou période) de rupture*** », nous désignons une date (ou une période) de transition vers une modification profonde et durable (plusieurs années) de l'assolement qui traduirait des changements de pratiques agricoles.

correspondant à des modifications de l'assolement. Pour analyser ces modifications d'assolement, un moyen de procéder consiste à segmenter la période d'étude en plusieurs sous-périodes, et analyser les régularités correspondant à des sous-périodes encadrant une date de rupture. Nous avons mis au point une méthode pour définir le nombre minimum d'états du HMM2 linéaire permettant de segmenter la période d'étude en sous-périodes disjointes (figure 3.4). Cette méthode consiste à tester des HMM2 linéaires ayant un nombre croissant d'états et de choisir le modèle qui permet la segmentation la plus satisfaisante de la période d'étude.

Pour les deux territoires d'étude, nous avons identifié des dates de rupture après lesquelles l'assolement connaît des modifications sur plusieurs années (figure 3.3). Pour analyser ces modifications d'assolement nous avons testé plusieurs longueurs de modèles linéaires. La figure 3.4(a) illustre le choix d'un HMM2 qui segmente la période d'étude en 6 sous-périodes dont deux encadrent des épisodes de sécheresses estivales qui ont influencé le raisonnement des agriculteurs dans le site d'étude de Chizé. Les régularités capturées par les états correspondant à ces deux sous-périodes ont été examinées afin de vérifier, à l'échelle du territoire agricole, la généricité de stratégies identifiées dans un échantillon réduit d'exploitations agricoles enquêtées (*cf.* chapitre 6 page 130). Un autre exemple illustrant le choix du modèle est donné dans la figure 3.4(b). Le HMM2 sélectionné segmente la période d'étude de manière biunivoque en 6 sous-périodes disjointes. Nous avons utilisé ce modèle pour spatialiser l'évolution de l'assolement dans le bassin versant du Yar (*cf.* section 3.2.2 page 59 et annexe A page 179).

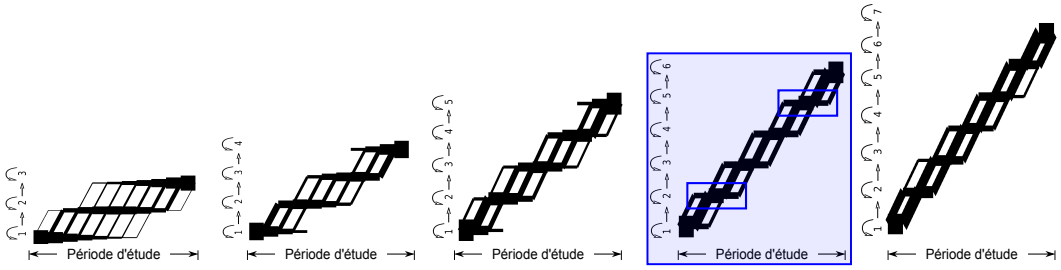


(a) **(b)**

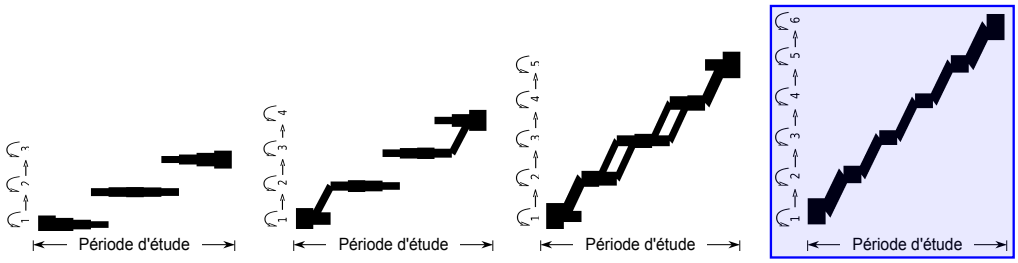
FIGURE 3.3 – Évolution de la fréquence des catégories d'OCS dans le site d'étude de Chizé (a) et des OCS dans le bassin versant du Yar (b) durant les périodes d'étude respectives. Pour le site d'étude de Chizé, les observations élémentaires sont des catégories d'OCS issues du regroupement des 45 OCS de la base de données en 11 catégories (*cf.* tableau 3.1). Pour le bassin versant du Yar les observations élémentaires sont les 6 OCS initiales de la base de données. L'année 2003 marque un moment de rupture où les Prairies et le Maïs cèdent la place aux Céréales dans ce bassin versant.

Recherche des régularités de transition entre les cultures

La recherche des régularités de transition entre les cultures vise à identifier les transitions entre les cultures, à estimer leurs fréquences et leur éventuelles évolutions. Cette analyse temporelle est fondée sur des observations élémentaires (OCS ou catégories d'OCS) pour lesquelles nous cherchons à identifier les transitions inter-annuelles. La recherche de ces régularités est réalisée avec un HMM2 ergodique (*cf.* section 2.3.2 page 39) qui compte autant d'états que d'observations élémentaires pour lesquelles il s'agit



(a)



(b)

FIGURE 3.4 – Recherche d'une segmentation satisfaisante de la période d'étude avec des HMM2 linéaires ayant un nombre croissant d'états. L'épaisseur des traits est proportionnelle à la probabilité de transition *a posteriori* entre deux états du HMM2. (a) La segmentation recherchée comporte au moins 2 segments temporels de plusieurs années ne se recouvrant pas et contenant le moins possible d'OCS indéterminées (correspondant aux parcelles non enquêtées) fréquentes au début de la période d'étude (*cf.* figure 2.3 page 20). Dans le HMM2 de 6 états, les états 2 et 5 identifient deux périodes disjointes. (b) La segmentation recherchée associe de manière biunivoque chaque état à un seul segment temporel de plusieurs années. Le modèle de 6 états segmente la période d'étude en 6 sous-périodes distinctes d'une durée moyenne de 2 ans.

d'estimer les probabilités de transitions. Lorsque le nombre des observations élémentaires est important, il est commode d'individualiser les observations d'intérêt dans des ***états de Dirac*** [4] et de définir un ***état de réserve*** [5] pour capturer le reste des observations. Les diagrammes de Markov obtenus à l'issue de cette analyse montrent les transitions inter-annuelles entre les observations élémentaires de manière proportionnelle à l'épaisseur des lignes (figure 3.5). La figure 3.5(a) montre que les transitions entre Tournesol, Blé et Colza sont les plus fréquentes dans le site d'étude de Chizé. Ce diagramme montre aussi par exemple que la monoculture du Maïs a considérablement régressé après l'année 2004. Ce résultat complète celui de la figure 3.3(a) qui montre une régression importante des surfaces en Maïs à partir de 2004. La figure 3.5(b) montre par exemple une diminution des transitions Maïs-Prairies contre une augmentation des transitions Maïs-Céréales, ce qui complète l'information apportée par la figure 3.3(b) où après l'année 2003, les surfaces en Prairies régressent au profit des Céréales : le maïs change de voisin temporel.

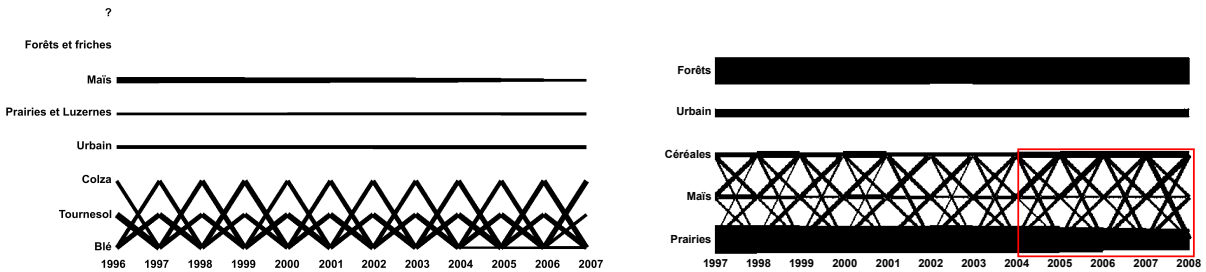


FIGURE 3.5 – Diagrammes de Markov montrant les transitions entre les cultures dans le site d'étude de Chizé (a) et dans le bassin versant du Yar (b). L'axe des abscisses représente la période d'étude. L'axe des ordonnées représente les états du HMM2 ergodique utilisé pour la fouille de données. Chaque état est associé à une seule observation (état de Dirac) sauf l'état « ? » (état de réserve) qui est associé à une distribution uniforme des observations. L'état de réserve est destiné à capturer les observations non spécifiées dans les états de Dirac. Les transitions diagonales représentent des transitions inter-annuelles d'OCS. Les lignes horizontales indiquent une stabilité inter-annuelle (monoculture). Par souci de visibilité, seules les transitions dont les fréquences sont supérieures à un seuil sont affichées. L'épaisseur des traits est proportionnelle à la probabilité de transition *a posteriori* entre deux états du HMM2.

Recherche des régularités des successions de cultures

Ce type de fouille temporelle cherche à identifier les principales successions de cultures, et à estimer leurs fréquences et leurs éventuelles évolutions. Cette analyse est fondée sur des successions d'observations élémentaires (OCS ou catégories d'OCS) dont la longueur est choisie selon les critères statistiques présentés au début de ce chapitre (*cf.* section 3.1.2 page 50). Nous avons recherché ces régularités temporelles pour le site de Chizé en choisissant comme observation : les successions sur 4 ans des catégories d'OCS que nous appellerons « successions » pour simplifier. Nous décrivons dans ce qui suit trois

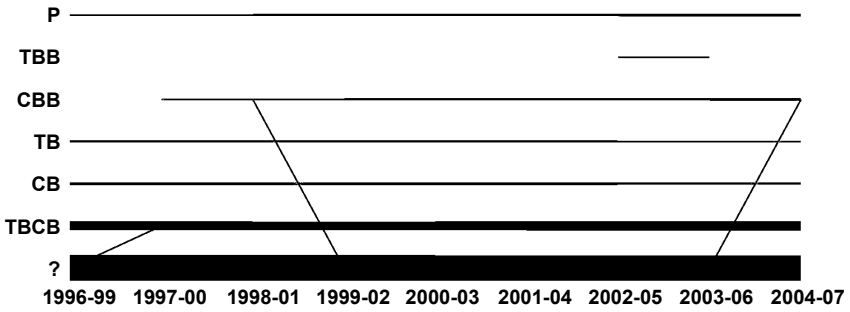
4. Un état de ***Dirac*** est associé à une distribution nulle sauf pour un nombre restreint d'observations.

5. L'***état de réserve*** est un état défini avec une distribution d'observations équiprobable avant l'apprentissage du modèle. Après l'apprentissage, cet état permet de capturer les observations non définies dans l'état de Dirac.

approches complémentaires que nous avons utilisées pour rechercher les régularités temporelles des successions.

La première approche recherche les successions majoritaires et leurs éventuelles évolutions. Elle utilise un HMM2 linéaire dont les états équiprobables sont initialisés avec des distributions uniformes des successions constituant l'univers du possible (environ 8 000 successions différentes, *cf.* section 3.1.2 page 50). Le nombre des états est défini suivant la méthode décrite précédemment (*cf.* section 3.2.1 page 53). Nous obtenons l'évolution des distributions des successions sur autant de sous-périodes que d'états définis dans le HMM2 linéaire.

La seconde approche estime la fréquence des rotations et leurs éventuelles évolutions. Elle s'appuie sur la connaissance (experte ou acquise par des fouilles préalables) des rotations culturales pratiquées dans le territoire d'étude. La recherche des régularités est réalisée avec un HMM2 ergodique (*cf.* section 2.3.2 page 39) dont chaque état est initialisé avec les successions correspondant à la rotation qui donne son nom à l'état. Par exemple, la rotation biennale Tournesol-Blé (TB) correspond aux 2 successions T-B-T-B et B-T-B-T, et la rotation triennale TBB correspond aux 3 successions T-B-B-T, B-B-T-B et B-T-B-B. L'état de réserve permet de capturer le reste des successions pour lesquelles aucune rotation n'a été identifiée. Cette approche présente, cependant, l'inconvénient de nécessiter la connaissance de l'ensemble ou de la majorité des rotations pratiquées dans le territoire agricole étudié. Sans cette information, l'état de réserve devient important (figure 3.6) et les résultats peu exploitables surtout pour les comparer aux résultats d'un autre territoire agricole ou pour les spatialiser.



T : Tournesol, B : Blé, C : Colza, P : Prairies et Luzernes, ? : État de réserve

FIGURE 3.6 – Diagramme de Markov montrant les principales rotations culturales identifiées dans le site d'étude de Chizé. L'axe des abscisses représente la période d'étude divisée en sous-périodes de 4 ans. L'axe des ordonnées représente les états du HMM2 ergodique utilisé pour la fouille de données. Les lignes horizontales indiquent une stabilité inter-annuelle des rotations. Par souci de visibilité, seules les transitions dont les fréquences sont supérieures à 2% sont affichées. L'épaisseur des traits est proportionnelle à la probabilité *a posteriori* des états du HMM2.

La troisième approche extrait des classes de successions. Elle est réalisée avec un HMM2 ergodique dont la topologie est donnée par la figure 3.7. Chaque état du HMM2 est initialisé par une distribution uniforme de successions correspondant aux motifs de recherche décrits dans le tableau 3.2. Les observations correspondant à un motif de recherche définissent une classe de successions. Les classes de successions sont notées S(X), X étant une catégorie d'OCS, et la classe S(X) se lit : classe des successions avec X. Chaque succession comporte au moins une occurrence de la catégorie d'OCS ayant donné son nom à la classe. Dans le territoire d'étude, le Blé (B), le Tournesol (T) et le Colza (C) sont le plus souvent intégrés

dans une même succession de 4 ans (p.ex. T-B-C-B, T-B-T-B, C-B-C-B, T-B-B-B, C-B-B-B). Une classe commune à ces trois cultures a permis d'obtenir des résultats plus cohérents qu'avec les classes séparées. Cette classe, notée S(B,T,C), regroupe les successions impliquant au moins l'une de ces trois cultures.

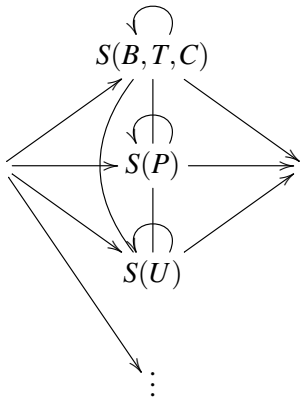


FIGURE 3.7 – Topologie du HMM2 utilisé pour la recherche des régularités des successions de cultures pour le site d'étude de Chizé. Chaque état représente une classe de successions S(X) où X est une catégorie d'OCS. T : Tournesol, B : Blé, C : Colza, U : Urbain, P : Prairies et Luzernes. Tous les états ne sont pas représentés.

TABLEAU 3.2 – Schéma général du motif de recherche utilisé pour initialiser les états du HMM2 de la figure 3.7. Ce motif permet de capturer les successions impliquant la catégorie d'OCS X au moins une fois parmi les 4 années de la succession. Le « ? » représente une quelconque catégorie d'OCS. L'ensemble des successions capturées dans un état constitue la classe des successions S(X). Par exemple S(P) est la classe des successions impliquant la catégorie « Prairies et Luzernes » au moins une année parmi les 4 années de la succession.

Année t	Année $t+1$	Année $t+2$	Année $t+3$
X	?	?	?
?	X	?	?
?	?	X	?
?	?	?	X

Les distributions uniformes des successions associées aux états du HMM2 sont réestimées avec l'algorithme Forward-Backward lors de l'apprentissage du HMM2 sur le corpus de données. Après l'apprentissage, ces distributions se présentent sous forme d'une liste de successions par état où chaque succession est associée à sa fréquence dans le corpus de données. Les résultats de cette première étape de l'analyse montre des classes de successions caractérisées par une monotonie marquée le long la période d'étude (1996-2007) suggérant l'intérêt de localiser ces régularités temporelles. Le diagramme de Markov et les distributions réestimées des classes de successions sont consultables dans le chapitre 4.

3.2.2 Spatialisation des régularités temporelles

La deuxième étape de la segmentation temporo-spatiale consiste à spatialiser les régularités temporelles, identifiées dans l'étape précédente, en patches caractérisés par des distributions homogènes de ces régularités. Cette étape de spatialisation est effectuée avec des HMM2 hiérarchiques (HHMM2) comme ceux décrits dans la figure 3.8. Nous décrivons dans cette section la spatialisation de deux types de régularités temporelles : l'évolution de l'assolement (*cf.* section 3.2.1) et les classes des successions (*cf.* section 3.2.1).

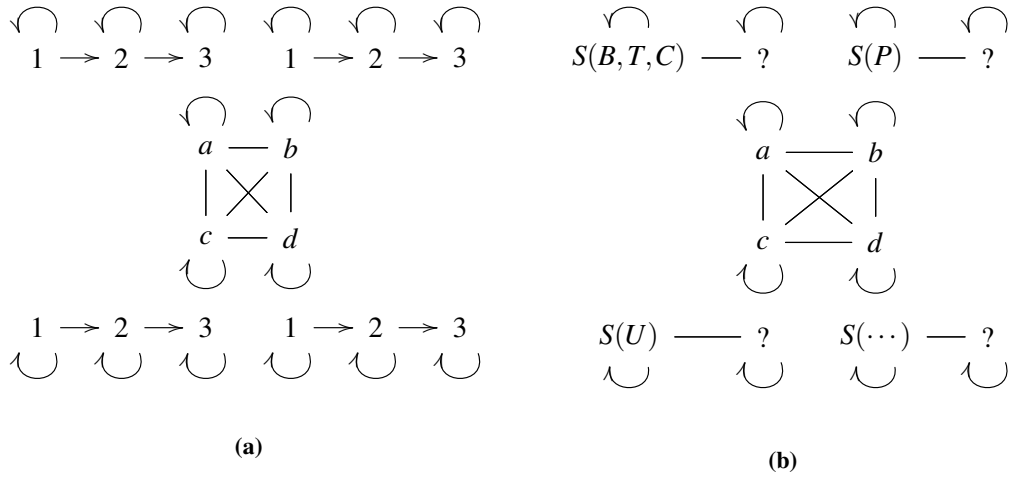


FIGURE 3.8 – Topologie des HHMM2 utilisés pour la segmentation temporo-spatiale des territoires d'étude. (a) HHMM2 constitué d'un HMM2 ergodique dont les états a, b, c et d sont associés à des HMM2 linaires dont les états sont 1, 2 et 3. (b) HHMM2 constitué d'un HMM2 ergodique dont les états a, b, c et d sont associés à des HMM2 ergodiques constitués de 2 états : un état de gabarit associé à une classe S(X) de successions avec la catégorie X d'OCS, et un état équiprobable associé à une distribution uniforme des successions. Tous les états ne sont pas représentés. Les traits non fléchés sont bidirectionnels.

Spatialisation de l'évolution de l'assolement

L'évolution de l'assolement a été spatialisée pour les deux territoires d'étude à l'aide de modèles hiérarchiques dont la topologie est donnée par la figure 3.8(a). Ces modèles sont construits par l'imbrication de deux types de modèles. Un HMM2 ergodique — appelé modèle maître — et des HMM2 linéaires associés aux super états du HMM2 maître. Les super états sont des états spatiaux destinés à capturer des patches caractérisés par les régularités temporelles que capturent les HMM2 linéaires. La segmentation spatiale recherchée doit aboutir à un nombre maximum de classes spatiales non redondantes. Pour ce faire, nous avons choisi le nombre des états spatiaux en testant plusieurs segmentations avec des HHMM2 au nombre croissant de super états. La redondance des classes est mesurée par la distance entre les distributions associées aux états, cette distance est évaluée à l'aide de la divergence (*cf.* équation 2.11). Le résultat de la segmentation est une carte de patches dont le contenu est caractérisé par l'évolution des assolements comme l'illustre la figure 3.9. Cette figure met en lumière une liaison stable par tranches de périodes entre Prairies, Céréales et Maïs. Cette liaison est ponctuellement rompue (2000) puis profondément modifiée (2003). D'autres segmentations de ce type sont rapportées dans le chapitre 6 (SCHALLER et al., 2011) et dans l'annexe A.

Spatialisation des S(X)

Les classes des successions ont été spatialisées pour le site d'étude de Chizé à l'aide du HHMM2 décrit dans la figure 3.8(b). Ce modèle est constitué d'un HMM2 maître ergodique dont les états spatiaux sont associés à d'autres HMM2 ergodiques constitués de 2 états : un état de gabarit associé à une classe S(X) de successions, et un état résiduel équiprobable associé à une distribution uniforme des successions. L'état

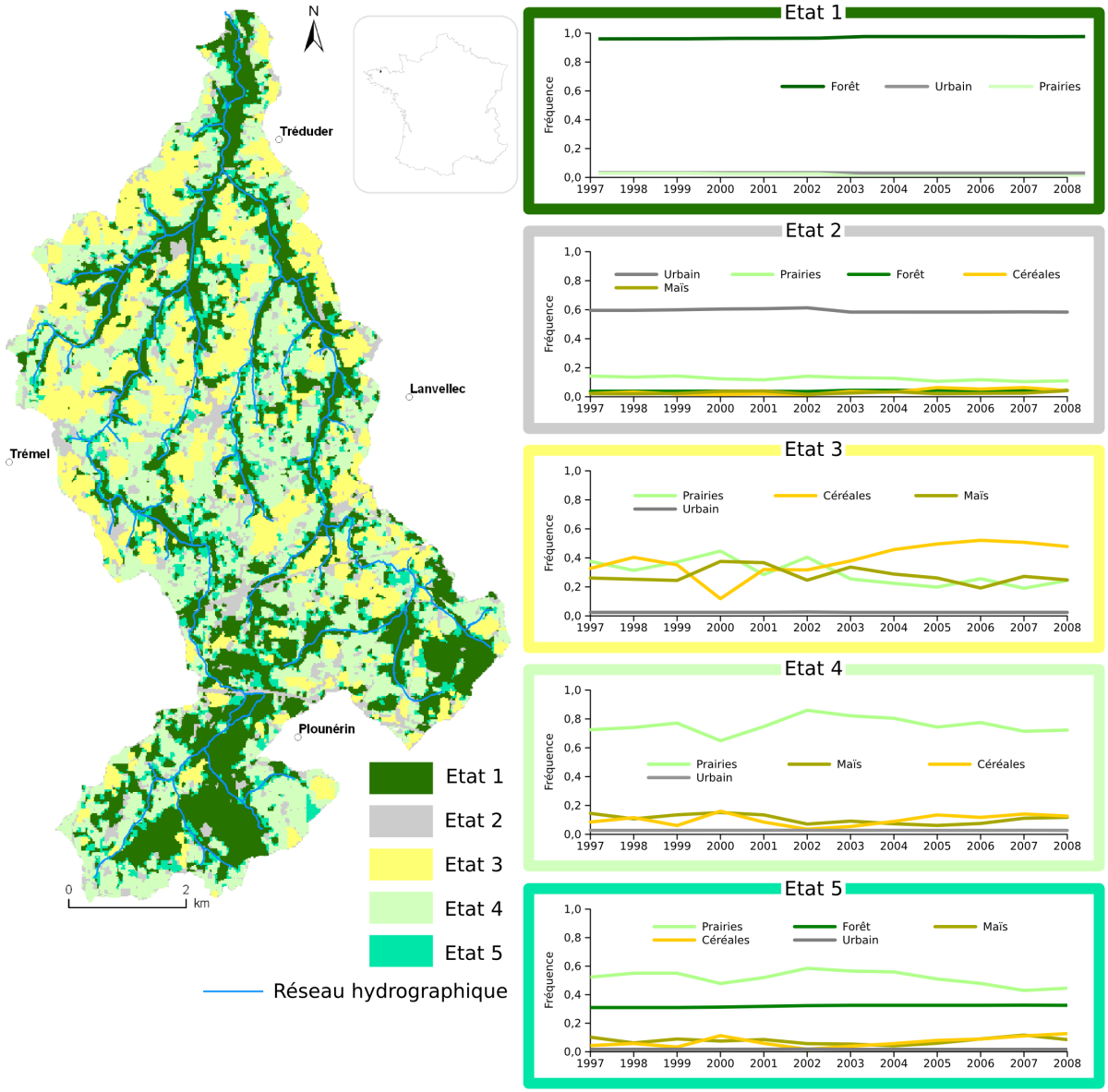


FIGURE 3.9 – Carte du bassin versant du Yar segmenté en patches d'évolution des OCS. Chaque unité cartographique correspond à un super état du HHMM2 qui a servi à la segmentation. Chaque super état est décrit par un diagramme d'évolution des distributions des fréquences des OCS. A partir de 2003, Prairies et Maïs régressent et sont remplacés par les Céréales dont la fréquence a nettement augmenté dans les patches appartenant à l'état 3.

résiduel est destiné à capturer les observations qui font exception au motif défini dans l'état de gabarit et de permettre l'obtention de patches « sans trous ». Nous verrons ensuite (*cf.* section 3.3.1 page 62) que cet état s'est avéré être également un état « de voisinage » capable de capturer les successions à la lisière des patches.

Le résultat de la segmentation est une carte de patches caractérisés par deux distributions de successions : une distribution caractéristique de l'état de gabarit qui décrit le contenu des patches de la classe spatiale correspondante et dont la proportion renseigne sur l'homogénéité spatiale des patches en termes des successions correspondant à la classe S(X). L'autre distribution caractérise l'état résiduel. Après apprentissage, nous avons recalculé les fréquences des successions appartenant aux deux distributions à partir des comptes des successions pour que l'ensemble des fréquences des successions soit égal à 1. La figure 3.10 présente le résultat de cette segmentation en décrivant les patches par la proportion des successions dans les états de gabarit. Une description plus détaillée de ces patches est donnée dans le chapitre 4.

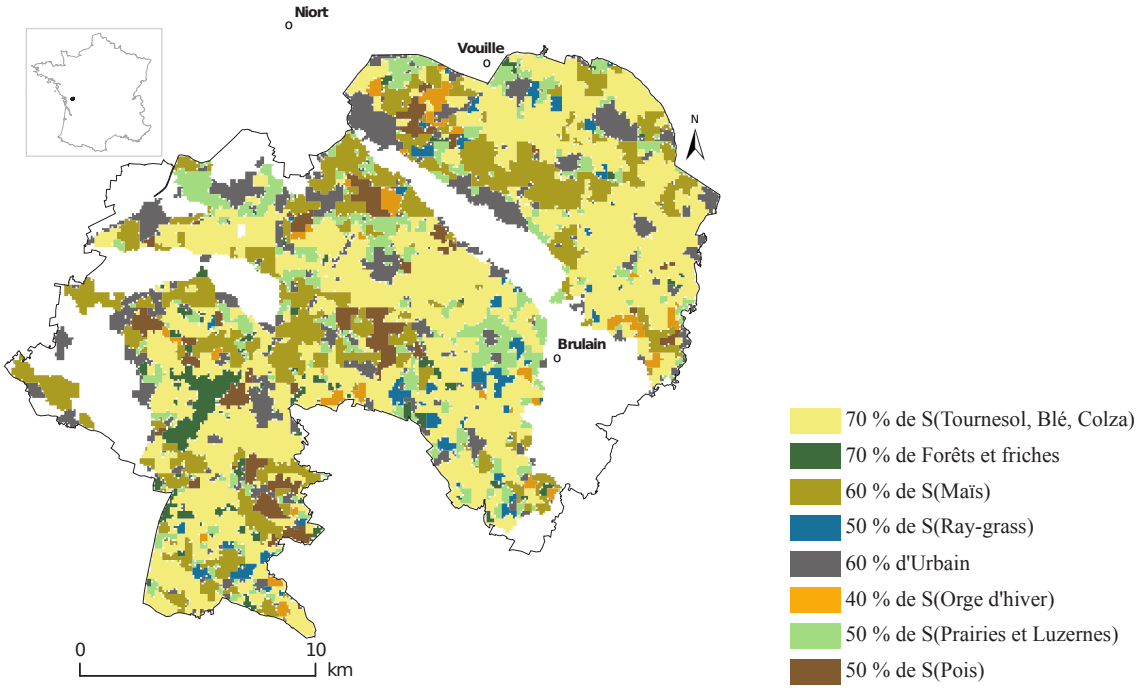


FIGURE 3.10 – Le site d'étude de Chizé décrit par des patches de successions. Chaque unité cartographique correspond à un état spatial du HHMM2 qui a servi à la segmentation. Ces unités cartographiques sont décrites par deux distributions de successions : une distribution associée à la classe S(X), et une autre distribution résiduelle associée à l'état équiprobable. Dans cet exemple, les unités cartographiques sont sommairement décrites par la proportion des successions dans les états de gabarit. Les zones blanches à l'intérieur du territoire d'étude n'ont pas été cartographiées car elles ont été insuffisamment enquêtées durant la période d'étude.

3.3 Fouille des voisinages des OCS et des successions d'OCS

Dans cette section nous présentons une méthode de fouille que nous avons développée pour décrire les voisinages des OCS et des successions d'OCS. Cette méthode est fondée sur deux approches complémentaires :

La première approche décrit les voisinages entre les patches de successions d'OCS. Cette approche est issue de la segmentation temporo-spatiale des successions et utilise une propriété de l'état résiduel qui permet de capturer les successions à la lisière des patches (*cf.* section 3.2).

La seconde approche décrit les voisinages d'OCS et de successions d'OCS entre les parcelles et les éventuelles évolutions de ces voisinages.

Nous décrivons dans cette section les grandes lignes de cette méthode que nous avons appliquée au site d'étude de Chizé.

3.3.1 Fouille du voisinage des patches

Nous avons vu précédemment (*cf.* section 3.2.1) que pour segmenter le site d'étude de Chizé en patches de successions, nous avons défini un état équiprobable initialement destiné à capturer les successions résiduelles minoritaires à l'intérieur des patches (chapitre 4 page 76). Cet état s'est avéré être également un état « de voisinage » capable de capturer les successions à la lisière des patches (MARI et al. (2010), chapitre 5). Afin d'illustrer cette propriété de l'état de voisinage, prenons un exemple de deux classes de successions S(P) (la classe des successions impliquant la catégorie « Prairies et Luzernes » au moins une année parmi les 4 années de la succession) et S(U) (la classe des successions impliquant la catégorie « Urbain ») dont les patches correspondants s'avoisinent régulièrement (*cf.* figure 3.10). Dans chacune de ces deux classes de patches, l'état de voisinage capture des successions appartenant à l'autre classe (tableau 3.3). L'état de voisinage de la classe Urbaine montre que les Prairies et luzernes sont les voisins les plus fréquents de l'Urbain. Inversement, l'Urbain est un voisin fréquent des Prairies et luzernes, mais les voisins les plus fréquents sont des successions du type S(B,T,C). Cette dissymétrie du voisinage est explicable par la tendance des Prairies et luzernes à former des couronnes autour de l'Urbain ce qui réduit ainsi le contact de l'Urbain avec les autres classes de patches. D'autre part, les Prairies et luzernes voisins d'un côté avec l'Urbain se trouvent ouverts au voisinage avec d'autres classes de patches et notamment avec la matrice agricole prédominante représentée par les successions du type S(B,T,C). La figure 3.11 illustre schématiquement cette interprétation.

L'état de voisinage capture les successions qui ne répondent pas à la définition des successions de l'état de gabarit et qui se trouvent au voisinage de ces successions. Nous avons vu que ces successions d'exception capturées dans l'état de voisinage peuvent se trouver à l'intérieur des patches. Dans ce cas l'état de voisinage joue un rôle de lissage spatial permettant d'obtenir des patches « sans trous ». Les successions d'exception se trouvent aussi à la lisière des patches et renseignent sur les successions majoritaires des patches voisins.

TABLEAU 3.3 – Distributions des successions dans deux unités cartographiques régulièrement voisines (*cf.* carte de la figure 3.10). Chaque couleur du tableau décrit l'unité cartographique correspondante à l'aide de 2 distributions : la classe S(X) des successions avec la catégorie X d'OCS qui décrit les patches par leurs contenus, et la distribution associée à l'état de voisinage qui décrit les patches par les successions à leurs lisières. La somme des fréquences des successions dans chaque couple de distributions (S(X), état de voisinage) égale 1. P : Prairies et luzernes U : Urbain. « ? » désigne l'information manquante dans le corpus de données correspondant aux parcelles non enquêtées. Seules les dix successions majoritaires par distribution sont affichées.

50% S(P)		50% État de voisinage		60% S(U)		40% État de voisinage	
P-P-P-P	0,20	T-B-T-B	0,03	U-U-U-U	0,60	P-P-P-P	0,03
Y-P-P-P	0,01	B-T-B-T	0,02	?-U-U-U	0,01	?-?-?-?	0,03
B-P-P-P	0,01	T-B-C-B	0,02	?-?-U-U	0,01	T-B-C-B	0,01
P-P-P-B	0,01	B-T-B-C	0,01	?-?-?-U	0,01	B-T-B-C	0,01
?-P-P-P	0,01	C-B-T-B	0,01	?-?-?-?	0,00	C-B-T-B	0,01
P-P-P-T	0,01	B-C-B-T	0,01	P-U-U-U	0,00	T-B-T-B	0,01
B-Y-P-P	0,01	B-T-B-B	0,01	U-U-U-B	0,00	B-C-B-T	0,01
T-P-P-P	0,00	U-U-U-U	0,01	B-U-U-U	0,00	B-U-U-U	0,01
P-P-T-B	0,00	B-B-T-B	0,01	?-P-U-U	0,00	U-B-U-U	0,01
T-B-P-P	0,00	C-B-C-B	0,01	T-U-U-U	0,00	B-T-B-T	0,00

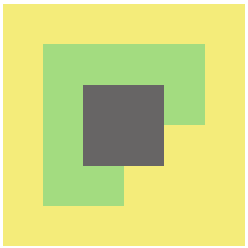


FIGURE 3.11 – Schéma expliquant la dissymétrie du voisinage entre les patches de l'Urbain et les patches des Prairies et luzernes qui ont tendance à entourer les zones urbaines. Le tout est entouré d'une matrice agricole ou prédominent les successions à base de Tournesol, Blé et Colza.

De plus, nous avons montré que l'état de voisinage permet de capturer les successions minoritaires au voisinage des patches. Nous avons testé cette propriété sur un corpus « chimère » construit à partir du corpus spatio-temporel des données d'OCS du site d'étude de Chizé. Ce corpus chimère a été préparé en remplaçant l'OCS de petits polygones ($< 0{,}5\,ha$) de l'Urbain par 3 OCS fictives (a, b, c) que nous avons répétées sur toute la période d'étude dans l'ordre de la succession « a-b-c ». L'ensemble des polygones associées à cette rotation représente environ 0.2% de la surface totale du territoire d'étude (figure 3.12). Alors que l'analyse temporelle ne permettait de déceler ni les OCS, ni la succession en raison de leur faible fréquence, l'état de voisinage a capturé les permutations circulaires de cette succession parmi la première dizaine des successions les plus fréquentes (tableau 3.4). Ce résultat pourrait être considéré comme une ébauche d'une méthode pour extraire des successions minoritaires intéressantes comme, par exemple, les successions innovantes. Mais afin de rendre cette procédure de fouille applicable à la recherche de successions innovantes, il faut d'abord définir des critères permettant de reconnaitre ces successions parmi les autres successions capturées par les états de voisinages.

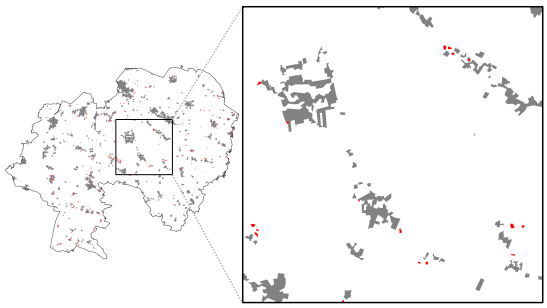


FIGURE 3.12 – Répartition spatiale de la succession a-b-c à la lisière des patches de l'Urbain. À gauche, vue sur l'ensemble du site d'étude de Chizé. À droite, zoom sur une partie du territoire d'étude.

TABLEAU 3.4 – Distributions des successions du corpus chimère dans l'état de gabarit et dans l'état de voisinage. Les trois permutations circulaires de la rotation abc apparaissent dans l'état de voisinage. P : Prairies et luzernes, U : Urbain, T : Tournesol, B : Blé, C : Colza, O : Orge d'hiver, A : Autres. « ? » désigne l'information manquante dans le corpus de données.

S(U)		État de voisinage	
U-U-U-U	0,606	P-P-P-P	0,029
?-?-?-?	0,021	?-?-?-?	0,008
?-U-U-U	0,012	T-B-C-B	0,007
?-?-U-U	0,005	B-T-B-C	0,006
?-?-?-U	0,005	C-B-T-B	0,006
P-U-U-U	0,002	T-B-T-B	0,006
?-U-U-U	0,001	O-U-U-U	0,005
U-U-U-B	0,001	U-O-U-U	0,005
A-U-U-U	0,001	B-C-B-T	0,005
P-P-U-U	0,001	a-b-c-a	0,005
B-U-U-U	0,001	b-c-a-b	0,005
B-A-U-U	0,000	c-a-b-c	0,005

3.3.2 Fouille du voisinage des parcelles

Dans cette approche nous cherchons à décrire les relations de voisinages d'OCS et de successions d'OCS entre les parcelles et à décrire les éventuelles évolutions de ces voisinages en vue de valider et de compléter les résultats de fouille du voisinage à l'échelle des patches (section 3.3.1). Nous analysons le voisinage des successions d'OCS en segmentant le territoire d'étude en zones compactes caractérisées par une homogénéité interne en termes des successions d'OCS voisines. L'évolution du voisinage des OCS

est, quant à elle, analysée pour l'ensemble du territoire d'étude en segmentant la période d'étude en sous-périodes encadrant une date de rupture [6].

Choix des observations pour la fouille du voisinage des parcelles

Pour l'analyse du voisinage des OCS et des successions d'OCS entre les parcelles, nous nous appuyons sur le quintuplé de voisins (*cf.* chapitre 2 page 37). Toutefois, les quintuplés génèrent une grande diversité d'observations qui alourdit l'apprentissage du modèle et complique l'interprétation des résultats. Lorsque le territoire d'étude est isotrope vis-à-vis de l'observation choisie, les quintuplés sont avantageusement remplacés par les cliques qui génèrent des diversités d'observations plus facilement interprétables [7]. Nous avons simplifié le quintuplé en clique (*cf.* encadré 3 page 40) après avoir vérifié l'isotropie du site d'étude de Chizé vis-à-vis des OCS en suivant la démarche décrite dans l'encadré 2 page 38. Nous avons retenu deux types d'observations : la clique temporelle d'OCS et la clique temporelle de successions d'OCS. La première est utilisée pour analyser l'évolution des voisinages des OCS. La seconde est utilisée pour analyser le degré d'attraction entre les successions d'OCS.

Que ce soit avec des quintuplés ou avec des cliques, l'étude du voisinage du premier ordre (c.à.d. des voisins immédiats) nécessite l'utilisation de fines résolutions d'échantillonnage afin que les résultats puissent être interprétés proportionnellement à la longueur des frontières entre les parcelles (ou les *parcelles élémentaires* [8] lorsque l'observation est une clique temporelle de successions d'OCS). Avec des résolutions plus grossières, les quintuplés (ou les cliques) atteignent plusieurs ordres de voisinages (voisins, voisins des voisins, ...) et les observations seront moins bien corrélées à la longueur des frontières.

Fouille des successions d'OCS dans les parcelles élémentaires voisines

Cette analyse cherche à simplifier la complexité apparente de l'OTAA en segmentant le territoire d'étude en patches caractérisés par des distributions de successions d'OCS voisines [9]. Les observations sont des cliques temporelles de successions d'OCS. La segmentation temporo-spatiale est réalisée à l'aide d'un HMM2 ergodique dont les états sont initialisés avec des distributions uniformes des observations.

La segmentation temporo-spatiale du site d'étude de Chizé a permis d'identifier des patches au sein desquels les successions d'OCS voisines manifestent de forts degrés d'attraction (LAZRAK et al. (2010) ; MARI et al. (2010), *cf.* chapitre 5 pages 115 et 122). La figure 3.13 résume les résultats obtenus et montre que les Prairies sont les voisins préférés de l'Urbain comme le montrait déjà la première approche de fouille du voisinage des patches de successions d'OCS décrite dans la section 3.3.1 (page 62). Ce résultat

6. La ***date (ou période) de rupture*** a été définie en section 3.2.1 page 53.

7. **Remarque :** Notons que le nombre de configurations des cliques est considérablement plus faible que celui des quintuplés. Alors que la diversité des cliques temporelles d'OCS ne dépasse pas n^2 observations différentes, n étant le nombre d'observations élémentaires, le nombre des quintuplés temporels d'OCS peut atteindre n^5 observations différentes. Pour les cliques et quintuplés temporels des successions d'OCS, cet écart dépend en plus de la longueur des successions.

8. La ***parcelle élémentaire*** a été définie en section 2.1.2 page 19

9. Rappelons que deux successions d'OCS sont dites ***voisines*** si elles occupent 2 parcelles élémentaires voisines (*cf.* encadré 3 page 40).

appuie notre hypothèse que l'état de réserve capture les successions à la lisière des patches [10]. D'autre part, la présente approche a permis d'extraire une nouvelle régularité que la première approche n'avait pas révélée : les patches contenant les rotations incluant le Tournesol, le Blé et le Colza (patches oranges et jaunes dans la carte) sont constitués de parcelles dans lesquelles l'OCS voisine sera l'OCS suivante. Ceci montre que l'organisation dans le temps de la mosaïque agricole implique une organisation dans l'espace. Ceci nous conduit à envisager la description de la mosaïque en termes de quartiers culturaux formés de parcelles dans lesquelles les mêmes rotations sont pratiquées afin de réduire la variabilité temporelle et spatiale de cette mosaïque.

Notons que la résolution utilisée dans cet exemple d'application étant plutôt grossière ($80m$), les voisinages explorés sont de différents ordres et les fréquences des observations dans les distributions caractérisant les patches ne sont pas proportionnelles aux longueurs des frontières entre les parcelles (*cf.* section 3.1.3 page 51).

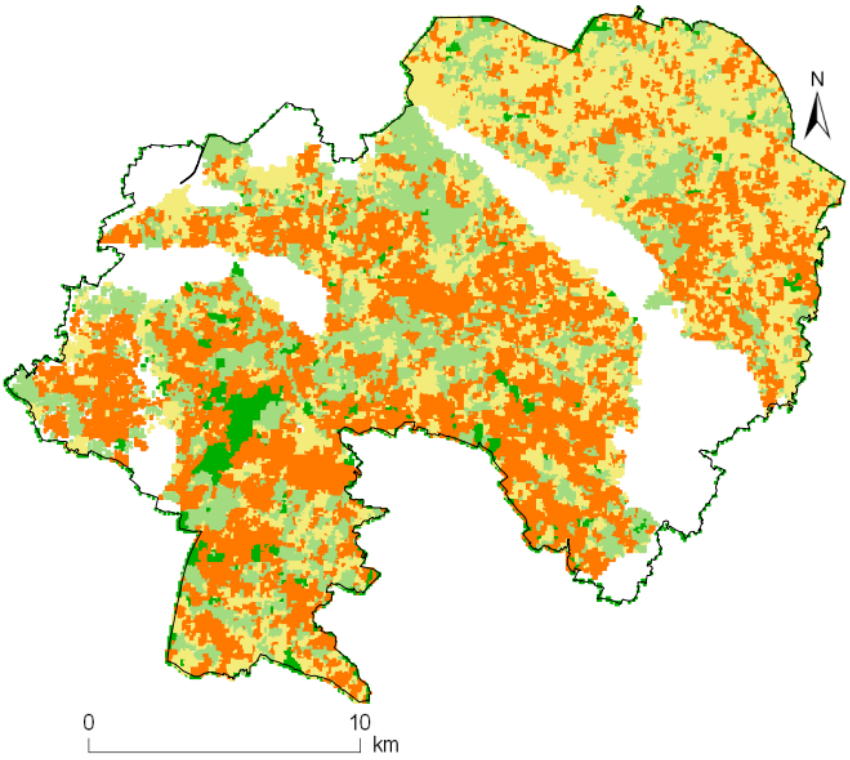
Fouille de l'évolution des OCS dans les parcelles voisines

Cette analyse cherche à décrire les relations de voisinage entre les OCS et à révéler d'éventuelles évolutions dans ces relations. Les observations sont des cliques temporelles d'OCS. La période d'étude est segmentée en sous-périodes disjointes avec un HMM2 linéaire. Les états sont initialisés avec des distributions uniformes de ces observations. À l'issue de la segmentation, chaque état correspond à une sous-période caractérisée par une distribution d'observations. Nous utilisons la PMI (*cf.* encadré 1 page 38) pour quantifier les relations de voisinages entre les OCS que nous interprétons respectivement par l'attraction ou la répulsion entre OCS selon que la valeur de la PMI est positive ou négative.

Nous avons cherché les régularités d'évolution du voisinage du Maïs et du Tournesol dans le site d'étude de Chizé. La segmentation temporelle a été réalisée avec le modèle linéaire sélectionné dans la figure 3.4(a) (page 55) en choisissant deux états qui correspondent à deux périodes disjointes encadrant une période de sécheresse estivale qui a influencé les logiques d'allocation de ces deux cultures dans un petit échantillon d'exploitations agricoles enquêtées.

La fouille de l'évolution du voisinage du Maïs et du Tournesol a révélé une attraction significativement croissante entre le Maïs et les Prairies et une répulsion significativement croissante entre le Tournesol et la Forêt. Ces résultats sont repris avec des éléments d'interprétation dans la section suivante (section 3.4.1 page 68) et dans nos articles LAZRAK et al. (2011) et SCHALLER et al. (2011) rapportés dans le chapitre 6 pages 130 et 142.

10. **Dans cette hypothèse**, nous avons supposé que les « états de voisinages » capturent les observations situées sur les parcelles voisines — occupées par des observations non définies dans l'état de gabarit — lorsque le parcours de Peano définit des trajectoires en « zigzag » de part et d'autre des frontières communes à ces parcelles. Cette hypothèse trouve sa justification dans le fait que le parcours de Peano ne reste pas suffisamment longtemps dans la parcelle voisine pour permettre au modèle de changer d'état spatial.



(a)

(X,X)-(X,X)	0,311	(P,P)-(P,P)	0,140	(C,C)-(B,B)	0,159	(T,T)-(B,B)	0,162
(F,F)-(F,F)	0,087	(M,M)-(M,M)	0,103	(B,B)-(C,C)	0,154	(U,U)-(U,U)	0,148
(?,?)-(?,B)	0,057	(S,S)-(B,B)	0,032	(B,B)-(B,B)	0,065	(B,B)-(T,T)	0,136
(?,?)-(M,M)	0,034	(U,P)-(U,P)	0,019	(B,B)-(O,O)	0,026	(M,M)-(B,B)	0,037
(?,?)-(P,P)	0,028	(P,U)-(P,U)	0,019	(B,B)-(A,A)	0,020	(B,B)-(M,M)	0,034
(X,?)-(X,?)	0,023	(M,M)-(P,P)	0,014	(B,C)-(C,B)	0,015	(T,B)-(B,T)	0,017
(?,?)-(T,T)	0,020	(T,P)-(B,P)	0,014	(C,B)-(B,C)	0,015	(B,T)-(T,B)	0,017
(?,?)-(U,U)	0,013	(P,T)-(P,B)	0,014	(B,B)-(Y,Y)	0,015	(B,B)-(P,P)	0,013
(U,U)-(U,U)	0,013	(B,P)-(T,P)	0,011	(O,O)-(C,C)	0,013	(O,O)-(T,T)	0,012
(?,?)-(C,C)	0,013	(P,B)-(P,T)	0,011	(T,C)-(B,B)	0,013	(T,T)-(T,T)	0,007
(X,U)-(X,U)	0,010	(M,M)-(T,T)	0,010	(C,T)-(B,B)	0,012	(T,U)-(B,U)	0,007
(U,F)-(U,F)	0,007	(C,P)-(B,P)	0,009	(B,T)-(C,B)	0,010	(U,T)-(U,B)	0,007
(F,U)-(F,U)	0,007	(Y,Y)-(P,P)	0,009	(T,B)-(B,C)	0,010	(B,U)-(T,U)	0,006
(?,?)-(B,P)	0,005	(P,C)-(P,B)	0,008	(Y,Y)-(B,B)	0,010	(U,B)-(U,T)	0,006
(?,?)-(P,B)	0,005	(P,B)-(P,C)	0,008	(B,B)-(T,C)	0,010	(T,B)-(B,B)	0,005
(X,C)-(X,B)	0,005	(B,P)-(C,P)	0,008	(B,B)-(C,T)	0,009	(C,T)-(B,B)	0,005
(F,C)-(F,B)	0,005	(M,P)-(M,P)	0,008	(A,A)-(B,B)	0,009	(B,T)-(B,B)	0,005
(X,T)-(X,B)	0,005	(B,B)-(?,?)	0,008	(C,B)-(B,T)	0,009	(T,C)-(B,B)	0,005
(?,?)-(B,T)	0,005	(T,T)-(M,M)	0,008	(B,C)-(T,B)	0,009	(T,T)-(O,O)	0,005
(C,F)-(B,F)	0,005	(M,M)-(A,A)	0,008	(C,B)-(B,B)	0,008	(B,B)-(T,B)	0,005
(?,?)-(T,B)	0,005	(P,M)-(P,M)	0,007	(B,B)-(C,B)	0,008	(P,P)-(M,M)	0,004

(b)

FIGURE 3.13 – Le site d'étude de Chizé décrit par des patches de cliques temporelles de successions biennales d'OCS (a). Chaque unité cartographique correspond à une distribution de ces cliques temporelles (b). Une clique temporelle représente 2 OCS successives sur 2 parcelles voisines qui se lit dans l'ordre suivant : $(S_t, V_t) - (S_{t+1}, V_{t+1})$, où S_t et V_t : sont respectivement les OCS de l'année t d'un site (S) et de son voisin (V) dans la courbe de Peano réduite, S_{t+1} et V_{t+1} : sont respectivement les OCS qui leur succèdent l'année suivante. X : Hors zone, F : Forêt et friches, B : Blé, T : Tournesol, C : Colza, U : Urbain, P : Prairies et Luzernes, M : Maïs, Y : Ray-grass, O : Orge d'hiver, A : Autres (*cf.* tableau 3.1 page 3.1). Les sites successifs sur la courbe fractale réduite résidant dans la même parcelle ne nous intéressent pas dans l'étude des voisinages entre parcelles. Nous nous intéressons dans les distributions aux observations (en rouge) qui vérifient : $S_t \neq V_t$ et(ou) $S_{t+1} \neq V_{t+1}$. Seules les 20 observations les plus fréquentes sont représentées. Les zones blanches à l'intérieur du territoire d'étude n'ont pas été cartographiées car elles ont été insuffisamment enquêtées durant la période d'étude.

3.4 Interactions avec les experts

Notre méthode de modélisation de l'OTAA est un cadre d'analyse où interagissent *l'analyste* qui fouille les données et fait émerger des régularités statistiques, et *l'expert* qui possède une certaine connaissance du territoire d'étude (*cf.* section 2.2.1 page 22). Notre interaction avec les experts s'inscrit dans une approche inductive où nous cherchons à établir des règles qui généralisent les comportements particuliers découverts grâce aux régularités extraites.

Nous avons interagi avec des experts de deux champs disciplinaires : avec des agronomes cherchant à comprendre les déterminants de l'OTAA en amont de notre objet de modélisation, et en aval avec des écologues cherchant à comprendre l'impact de l'OTAA sur la biodiversité (figure 3.14). Les connaissances (ou les hypothèses) de l'expert sont sous forme de règles logiques identifiées par des observations localisées (enquêtes à l'échelle de l'exploitation agricole, expérimentations, ...). En s'appuyant sur ces règles, l'expert interprète les régularités issues de fouilles de données exploratoires dont l'objectif est d'amorcer les interactions entre l'analyste et l'expert. Réciproquement, l'expert peut proposer à l'analyste de réaliser des fouilles ciblées afin de valider, à l'échelle du territoire d'étude, des connaissances acquises à des échelles plus locales. Nous présentons dans cette section deux expériences d'interactions avec ces experts.

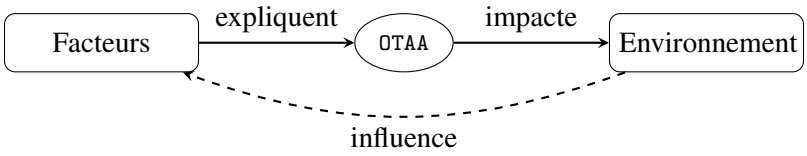


FIGURE 3.14 – Domaines disciplinaires des experts intervenant dans la modélisation de l'OTAA. L'OTAA est déterminée par des facteurs d'origines diverses (biophysiques, économiques, politiques, sociales, ...). Nous avons interagi avec des experts agronomes qui intègrent ces facteurs à travers l'identification de règles de décisions prises par les agriculteurs moyennant des enquêtes dans leurs exploitations. L'OTAA exerce des impactes sur l'environnement. Nous avons interagi avec des experts écologues qui étudient ces impacts sur la biodiversité.

3.4.1 L'expérience de fouille avec les agronomes

Nous avons articulé notre approche de fouille de données à l'échelle du territoire agricole étudié avec une approche d'enquêtes auprès d'un petit échantillon d'exploitations agricoles. La fouille de données contribue de deux façons à alimenter nos interactions avec les agronomes du terrain :

- La fouille de données permet d'identifier des régularités statistiques susceptibles de déboucher sur des hypothèses de règles de localisation qui peuvent être vérifiées par des enquêtes au niveau des exploitations agricoles.
- D'autre part, la fouille de données peut chercher les traces, dans la mosaïque du territoire agricole, de règles de décisions identifiées au niveau de l'exploitation agricole.

Nous avons testé ce cadre d'analyse dans le site d'étude de Chizé pour expliquer des régularités statistiques issues de fouilles concernant la culture du maïs, et pour vérifier la généricité, à l'échelle du territoire, de règles de décision d'agriculteurs enquêtés concernant la culture du tournesol.

Explication des régularités statistiques : illustration avec l'exemple du Maïs Une fouille exploratoire sur les dynamiques de voisinages des cultures montre que le maïs est devenu plus fréquent à proximité de prairies et moins fréquent à proximité des cultures commerciales (tournesol, blé et colza). Les résultats des enquêtes sur la gestion du maïs a donné aux experts quelques éléments explicatifs. Les agriculteurs ont dû adapter leurs stratégies au contexte local de l'interdiction de l'irrigation. Ces stratégies dépendent de leur système d'exploitation et de l'accès aux sols profonds. Quand ils ont accès à des sols profonds sur leur territoire d'exploitation, le maïs est cultivé en monoculture au voisinage des prairies situées traditionnellement dans les sols profonds et humides. Lorsque les agriculteurs n'ont pas accès à des sols profonds, leurs choix de cultures dépendent en grande partie du système d'exploitation, des besoins fourragers annuels et de la capacité d'irrigation (modulée par un risque de restriction). Par conséquent, le maïs a tendance à diminuer dans les exploitations de grandes cultures où il n'est pas une culture prioritaire, alors qu'il est maintenu dans les exploitations d'élevage avec accès à l'irrigation. En parallèle, la production du maïs irrigué étant devenue de plus en plus risquée, les éleveurs ont tendance à étendre les prairies pour compléter la production du fourrage. Le maïs est donc essentiellement maintenu sur les exploitations d'élevage avec des zones de prairies importantes sur leur territoire agricole. Toutes ces décisions des agriculteurs sont compatibles avec le fait que le maïs soit statistiquement de plus en plus proche des prairies et explique cette régularité au niveau du territoire agricole. C'est ainsi que des régularités à l'échelle du territoire agricole peuvent être expliquées par des décisions d'agriculteurs prises à l'échelle de l'exploitation agricole grâce à l'identification des règles de décision génériques.

Vérification de la généricité de règles de décision : illustration avec l'exemple du Tournesol En parallèle avec le premier exemple, des règles de décision individuelles identifiées par des enquêtes à l'échelle de l'exploitation agricole peuvent être validées à l'échelle du territoire agricole à travers l'identification de régularités d'OCS qui vont dans le sens des changements produits par la mise en œuvre de ces règles de décision. Nous illustrons ce cas par un exemple concernant la culture du tournesol dans le site d'étude de Chizé. Une règle de décision commune à plusieurs exploitations enquêtées consistait à diminuer la superficie consacrée à la culture du tournesol. Ce choix était justifié par des sécheresses estivales plus fréquentes qui affectent les rendements de tournesol, et aussi par la réforme de la Politique Agricole Commune (PAC). En conséquence, la plupart des agriculteurs ont décidé de réduire la superficie de culture consacrée au tournesol : ils ont arrêté de le cultiver à proximité des forêts en raison des dommages fréquents causés par les ravageurs. Cette règle a été validée par des fouilles à l'échelle du territoire agricole qui ont montré que la fréquence de tournesol a diminué globalement à l'échelle du territoire et particulièrement au voisinage des forêts et que la répulsion entre le tournesol et la forêt est devenue plus forte.

3.4.2 L'expérience de fouille avec les écologues

Nous avons tenté d'articuler notre approche de fouille de données avec des faits de biodiversité dans le site d'étude de Chizé à travers l'exemple du busard cendré (*Circus pygargus*) qui est un rapace se nourrissant de micro-mammifères (type campagnols essentiellement) et nichant principalement dans les champs de blé et de ray-grass qui offrent les hauteurs de végétations recherchées par cet oiseau migrateur à son arrivée dans la région. Nous disposions d'une base de données spatio-temporelle localisant les nids de busards annuellement et exhaustivement sur le site d'étude. Nous avons utilisé cette base de données pour mieux comprendre la contribution de l'OTAA dans la distribution des nids de busards. Dans un premier temps, nous avons présenté des résultats de fouille aux écologues qui nous ont aidé à mieux les comprendre. Dans un second temps, les écologues nous ont aidé à paramétrer une fouille centrée « animal » et à interpréter les résultats.

En vue de contribuer à l'identification des facteurs structurant la distribution spatiale des nids de busards dans le site d'étude de Chizé, nous avons croisé la localisation des nids avec les patches des classes de successions. Nous avons ensuite calculé les densités des nids par classe de patches. Comme nous l'attendions, ces densités sont les plus importantes dans les patches où prédominent les cultures de blé et (ou) de ray-grass (tab. 4, LAZRAK et al., 2009, *cf.* Chapitre 4). Cependant, la répartition hétérogène des nids au sein de ces patches renseigne sur l'existence d'autres facteurs structurants (fig. 6, LAZRAK et al., 2009, *cf.* Chapitre 4). Nous avons présenté ce résultat aux écologues qui l'ont partiellement expliqué par le comportement semi-colonial de cette espèce.

Les écologues nous ont ensuite orienté dans une fouille centrée « animal » que nous avons réalisée avec des cliques temporelles d'observations *composites*. Une observation est constituée d'une succession biennale d'OCS dans une parcelle élémentaire, associée à l'information de présence ou d'absence d'un nid de busard dans cette parcelle [11]. Cette fouille a permis d'identifier une classe spatiale qui a capturé la quasi totalité des parcelles ayant hébergé des nids de busards (figure 3.15). La distribution des observations dans cette classe est caractérisée par une proportion importante de luzernes : ce sont des patches où prédominent les parcelles cultivées avec des successions à base de blé, de tournesol et de colza et où prédominent aussi des luzernes pluriannuelles. Cette régularité est expliquée par les écologues par le rôle de « source » que jouent les éléments pérennes du paysage en plaine céréalière intensive dans la dynamique des métapopulations [12] du campagnol des champs. Les perturbations annuelles sur les surfaces cultivées entraînent l'extinction locale (à l'échelle de la parcelle) de ces populations, dont la persistance ne peut être assurée que par une dynamique de colonisation et de dispersion de ces micro-mammifères. Ce ré-

11. **Quelques précisions méthodologiques :** L'information ponctuelle « localisation des nids » étant relativement très faible comparée à l'information ponctuelle « OCS de l'ensemble des points de la grille de l'échantillonnage », la première avait tendance à « se noyer » dans la seconde. Nous avons résolu ce problème en considérant le nid comme une occupation de la parcelle et en considérant les parcelles se situant sur une zone tampon d'un rayon de $500m$ — défini par les écologues — autour du nid comme des parcelles occupées par ce nid. Afin de réduire le nombre de points de la grille d'échantillonnage, nous avons réduit les motifs de la fractale de Peano se trouvant à l'intérieur d'une parcelle élémentaire en les agglomérant en un point au centre de gravité de ce motif (*cf.* section 2.2.5 page 26). Notons enfin que la réduction de la fractale conduit à une normalisation des surfaces. Cette normalisation des surfaces a permis d'obtenir des résultats mieux contrastés comme si la notion de distance pour le busard n'était pas la même que la nôtre !

12. ***Une métapopulation*** est un groupe de populations d'individus d'une même espèce, séparées spatialement (ou temporellement) et qui réagissent réciproquement à un niveau quelconque (source : http ://fr.wikipedia.org/wiki/Métapopulation).

sultat conforte une hypothèse de travail forte des écologues sur le rôle clé du maintien d'un minimum de surfaces à végétation pluriannuelle au sein du paysage pour le maintien de la biodiversité.

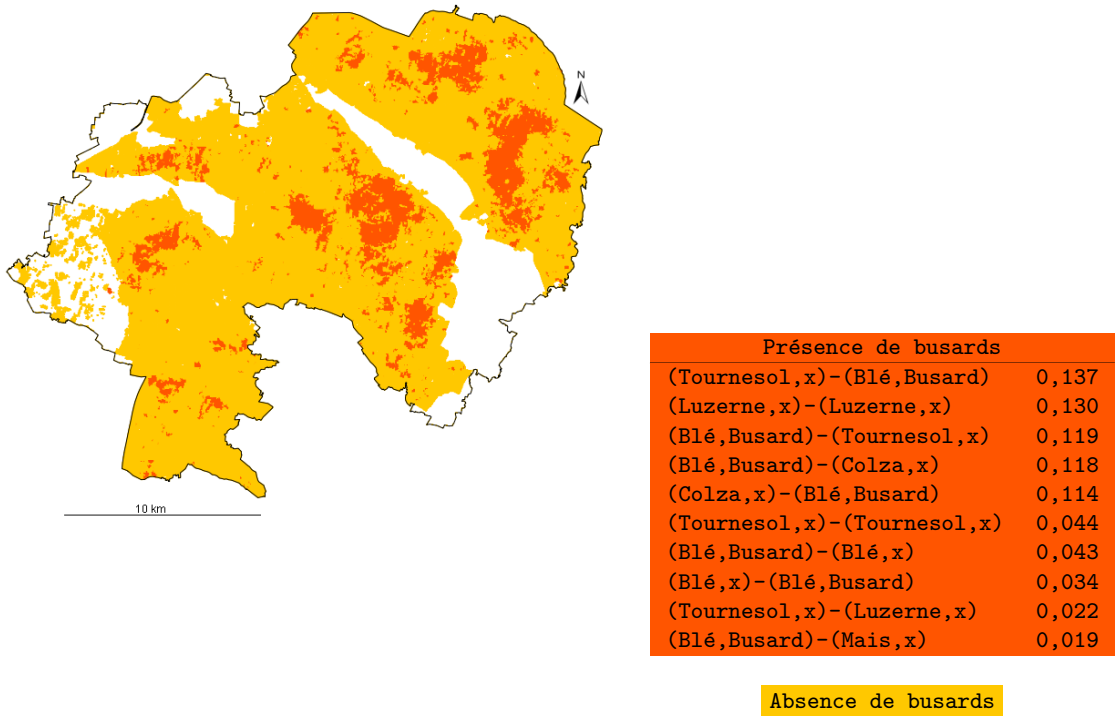


FIGURE 3.15 – Caractérisation des patches de distribution des nids de busards. La fouille a été réalisée avec des cliques temporelles d'observations composites constituées de l'OCS d'une parcelle élémentaire occupée ou non par un nid de busard sur deux années successives. L'absence du nid est notée « x ». Seules les 10 observations les plus fréquentes sont représentées. Les zones blanches à l'intérieur du territoire d'étude n'ont pas été cartographiées car elles ont été insuffisamment enquêtées durant la période d'étude.

3.5 Conclusion

Nous avons présenté dans ce chapitre une synthèse de la méthode de modélisation de l'OTAA développée dans le cadre de cette thèse. Cette méthode commence par un paramétrage préalable du corpus de données qui consiste à définir les observations élémentaires, la résolution de l'échantillonnage et la longueur des successions d'OCS. Ensuite, une étape exploratoire de fouilles temporelles permet d'identifier des régularités temporelles que nous spatialisons sous forme de cartes temporo-spatiales. Ce premier volet de la méthode aboutit à la description de l'OTAA sous forme de patches caractérisés par leurs contenus. Un deuxième volet de la méthode focalise sur les relations de voisinages dans lequel nous traitons deux niveaux de voisinages : le voisinage entre les patches et le voisinage entre les parcelles. L'analyse des voisinages à ces deux niveaux d'organisations a permis de vérifier l'existence de quartiers culturaux au sein desquels la variabilité spatio-temporelle des cultures n'est qu'apparente car elle concerne un même système de culture. Le troisième et dernier volet décrit nos interactions avec des experts en agronomie et en écologie. Ces interactions nous ont permis d'expliquer des régularités identifiées par des fouilles

exploratoires, et de construire ensemble des modèles de fouilles plus ciblés dans l'objectif de conforter leurs hypothèses.

La méthode ici présentée décrit d'une manière simplifiée, la complexité apparente de l'OTAA, et permet de faire une synthèse des dynamiques temporelles réelles sous forme de cartes de successions. Nous faisons l'hypothèse que la modélisation de l'OTAA par une image de successions est plus pertinente que l'approche classique consistant à représenter les dynamiques agricoles par l'analyse d'une succession d'images. Nous avons présenté dans ce chapitre des aspects aboutis de la méthode, mais aussi des ébauches de méthodes comme la fouille composite centrée « animal » qui présente l'avantage de combiner des faits agricoles avec des faits de biodiversité sans nécessiter des connaissances poussées sur le comportement de l'animal étudié. Une autre ébauche de méthode potentiellement intéressante est la recherche des successions innovantes qui s'appuie sur l'utilisation de l'état de voisinage, mais cette méthode nécessite d'être informée par des critères simples et pertinents décrivant les successions innovantes recherchées.

Les 3 chapitres suivants reprennent et complètent la présentation des aspects aboutis de notre méthode de modélisation :

Le chapitre 4 présente notre procédure de segmentation temporo-spatiale appliquée aux deux territoires d'études et déclinée dans trois articles.

Le chapitre 5 présente notre procédure de fouille des voisinages.

Le chapitre 6 présente deux articles illustrant nos interactions avec des experts agronomes dans un cadre d'analyse articulant deux approches complémentaires : la modélisation des régularités stochastiques sur les dynamiques de voisinage des OCS, et la modélisation des règles de décisions d'agriculteurs identifiées par enquêtes.

3.6 Références

LAZRAK, EG, M BENOÎT et J-F MARI (2010). « Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales ». Dans : *International Conference on Integrative Landscape Modelling*. Montpellier, France.

LAZRAK, EG, J-F MARI et M BENOÎT (2009). « Landscape regularity modelling for environmental challenges in agriculture ». Dans : *Landscape Ecology* 25.2, p. 169–183.

LAZRAK, EG, N SCHALLER et J-F MARI (2011). « Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol ». Français. Dans : *Atelier en marge d'EGC 2011*. Brest, France.

MARI, J-F, EG LAZRAK et M BENOÎT (2010). « Fouille de paysages agricoles: analyse des voisinages des successions d'occupation du sol ». Dans : *Colloque RTE (Raisonnement sur le Temps et l'Espace) en marge de RFIA 2010*. Sous la dir. de F LE BER, G LIGOZAT, O PAPINI et M BOUZID. Caen, France.

SCHALLER, N, EG LAZRAK, P MARTIN, J-F MARI, C AUBRY et M BENOÎT (2011). « Combining farmers' decision rules and landscape stochastic regularities for landscape modelling ». Anglais. Dans : *Landscape Ecology*.

Chapitre 4

Segmentation temporo-spatiale du territoire agricole

Dans ce chapitre nous présentons notre procédure de fouille de données permettant la segmentation temporo-spatiale d'un territoire agricole de dimension régionale à travers l'extraction de régularités temporelles puis leur spatialisation sous forme d'une carte où les unités cartographiques sont décrites par les distributions probabilistes de ces régularités. Cette procédure de fouille de données est illustrée à travers trois articles :

Le premier article (LAZRAK et al., 2009) présente la procédure de fouille de données appliquée au site d'étude de Chizé. Pour ce cas d'étude, la fouille de données a porté sur un corpus d'OCS construit par des relevés de terrain effectués le long de la période d'étude par des enquêteurs du CNRS à Chizé. La fouille de données a permis l'extraction des successions dominantes et leur spatialisation sous forme d'une carte de patches de successions.

Le deuxième article (LAZRAK et al., 2009), adressé à un public d'informaticiens, reprend les principales étapes de la procédure de fouille temporo-spatiale sans développer les aspects agronomiques.

Le troisième article, en cours d'élaboration, présente un test de généricité de la procédure de fouille initialement développée pour le cas du site d'étude de Chizé. Ce test de généricité a permis d'étendre l'utilisation de cette procédure de fouille pour une série d'assolements annuels construits à partir d'images satellites. La segmentation de ce corpus de données avec des modèles de Markov cachés a permis de décrire le Bassin Versant du Yar en patches d'évolution des assolements annuels. Cette modélisation a permis de révéler des dynamiques jusqu'alors cachées des systèmes de cultures pratiqués dans ce territoire agricole.

Landscape Ecol
DOI 10.1007/s10980-009-9399-8

RESEARCH ARTICLE

Landscape regularity modelling for environmental challenges in agriculture

El Ghali Lazrak · Jean-François Mari · Marc Benoît

Received: 18 November 2008 / Accepted: 17 August 2009

Abstract In agricultural landscapes, methods to identify and describe meaningful landscape patterns play an important role to understand the interaction between landscape organization and ecological processes. We propose an innovative stochastic modelling method of agricultural landscape organization where the temporal regularities in land-use are first identified through recognized Land-Use Successions before locating these successions in landscapes. These time–space regularities within landscapes are extracted using a new data mining method based on Hidden Markov Models. We applied this methodological proposal to the Niort Plain (West of France). We built a temporo-spatial analysis for this case study through spatially explicit analysis of Land-Use Succession dynamics. Implications and perspectives of such an approach, which links together the temporal and the spatial dimensions of the agricultural organization, are discussed by assessing the relationship between the agricultural landscape patterns defined using this approach and ecological data through an illustrative example of bird nests.

E. G. Lazrak (✉) · M. Benoît
INRA, UR 055 SAD ASTER, 88500 Mirecourt, France
e-mail: lazrak@mirecourt.inra.fr

J.-F. Mari
UMR 7503 LORIA, U. Nancy 2, 54506 Vandœuvre-lès-Nancy, France

Keywords Cropping system · Land-use changes · Temporo-spatial analysis · Data mining · Hidden Markov Model (HMM) · Hierarchical Hidden Markov Model (HHMM) · Landscape agronomy · Agricultural patterns · Landscape ecology · Poitou-Charentes (France)

Introduction

Agroecosystems are the major mode of land-use both at French (52%) and European (42%) levels. Since 1962, under the influence of the Common Agricultural Policy, agricultural production has intensified and its effects on biodiversity are no longer a matter of debate (Donald et al. 2001; Robinson and Sutherland 2002). Biodiversity conservation and restoration have become a social necessity and a political goal. Yet the practical means to achieve them in agroecosystems are still to be developed (Turner 2005; Butler et al. 2007). To understand the interaction between landscape organization and ecological processes, one needs to identify and quantify landscape patterns in meaningful ways (Turner 1990).

In agricultural landscapes, land-uses are heterogeneously distributed among different polygons (agricultural parcels). They also display dynamic patterns as a result of crop cycles, agricultural practices and other driving forces of land-use changes.

Recent studies (Retho et al. 2008) have investigated the relationship of restricted areas of agricultural

Published online: 16 September 2009

landscape with the diversity of animal species. On wider areas the complexity of agricultural landscapes needs to be simplified before investigating the relationships between a set of landscape indices (predictor variables) and ecological variables. This complexity of agricultural landscapes is generally simplified by using coarse agricultural land-use classes (e.g., Poudevigne and Alard 1997; Donald et al. 2001; Fuller et al. 2005). In this paper, we introduce a method which takes into account greater agricultural knowledge to identify and describe agricultural patterns. This description of agricultural patterns could improve the assessment of biodiversity in agricultural ecosystems at small geographical scales, and the assessment of water resource degradation in agricultural regions (Mignolet et al. 2007).

Agricultural biodiversity is related to soil, climate and cropping system interactions (Forman and Godron 1986). We focused on the factors managed and driven by farmers. For us, they correspond to cropping systems, which were categorized by Sebillotte (1974) into two components: (1) crop successions, seen as the ordered sequences of soil occupancy in each field and (2) technical sequences, defined as the ordered sequences of cultural practices on a crop for production.

This research falls within the realm of landscape agronomy, the aim of which is to study the organization of farming activities on small geographical scales (Benoît et al. 2007). Its scope is at the conjunction where technical farming activities in the fields, the influence of EU and world regulations impact on agriculture. We now see an increasing attention being given to local issues as result of farming activities and environmental preservation in particular. Landscape agronomy as an emerging discipline combines the concepts and methods of both (1) geographers and (2) agronomists (Benoît and Papy 1997), by respectively combining the following:

1. multi-scale modelling for land-use changes and for the investigation into territorial consistencies (Veldkamp and Fresco 1997; De Koning et al. 1999); and
2. analytical methods to describe the reasoning behind the way regional agricultural systems work. These methods notably rely on the construction and then on the spatialization of farming classifications (Perrot 1990; Landais 1996; Mignolet and Benoît 2001; Leisz et al. 2005; Mignolet et al. 2007).

Taking the production system into account as an organization level of farming activity makes it possible to place it in its political and socio-economic context. The production system is seen as the combination of soil, workforce and working methods used to produce crops and breed animals (Reboul 1976). We put forward the hypothesis that production systems determine landscapes and that the resulting agricultural landscapes can be described by studying the spatial organization of cropping systems. We modeled the farming activity by using the first component of cropping systems, i.e., crop successions.

The way a farmer organizes his territory is a time and spatial process. The land-use category of a given site depends upon the land-use categories of the neighborhood. For example, grasslands are mainly located in areas close to villages, whereas maize fields are usually far away from forests. The Markov Random Field (MRF) is an elegant mathematical model to take into account the uncertainty of locations in the vicinity of a given place. This model clusters the territory into patches where the distribution of land-use categories follows a certain probability law. On the other hand, the land-use category at time t (the current year is the usually admitted time slot unit) for a field depends also upon its former category at time $t-1$, $t-2$, etc. Since the late nineteenth century, plant successions have been studied on vegetation dynamics of natural ecosystems as reviewed by Glenn-Lewin and van der Maarel (1992). Among several approaches for studying vegetation dynamics, statistical models, especially Markov models, have already proved their usefulness (Usher 1992; Castellazzi et al. 2008). However, the precision of Markov models depends upon the quality of parameter estimation (Peet et al. 1992). The parameters of first-order Markov models can be tuned with the help of experts (Castellazzi et al. 2008). They can also be automatically estimated by means of algorithms such as the Baum–Welch algorithm (Welch 2003) when dealing with Hidden Markov Models (HMMs). In addition, compared to a first-order HMM, a higher-order HMM can adequately assign a probability to longer successions of land-use categories and reveal some temporal patterns.

In this paper, we propose a unified Markovian framework to:

1. represent both spatial and temporal dependencies in sites, and;
2. cluster a territory into patches where the successions of land-use categories are drawn by a higher-order Markov process.

Our main objective is to develop a generic method aimed at describing agricultural landscapes through cropping system patterns. As Moonen and Bàrberi (2008) emphasize the importance of territorial organization as a determinant of functional biodiversity, two applications of these patterns are:

1. to help understand environmental and natural processes in crop regions in Europe in relation with the organization and dynamics of agricultural landscapes, and;
2. to develop a knowledge of agricultural dynamics, which may facilitate political decision-making and forecasting.

After presenting the characteristics of the landscape we studied, we will give a brief theoretical background of our research procedure. Then, we explain this procedure in two stages. The first stage aims at modelling the diversity of the cropping systems, independently from their location within the landscape. The second stage is focused on the localization of these cropping systems that have been revealed in the previous stage. We propose to build a temporo-spatial analysis through spatial analysis of crop dynamics. This approach links together the temporal and the spatial dimensions of the agricultural organization. Its implications and perspectives are discussed in an illustrative example drawn from the Niort Plain by assessing the relationship between the agricultural landscape patterns and the distribution of Montagu's Harrier nests.

Methods

Study area

We apply our landscape regularity modelling method on one grain-growing area in France: the Niort Plain situated within the plain of Niort-Brioux, south of Deux-Sèvres in Poitou-Charentes region (46.2°N, 0.4°W).

Since 1994, the Chizé Centre for Biological Studies (CNRS[1] unit) has been conducting a research program on the impact of farm management strategies on biodiversity in grain-growing plains. This program focuses on the Niort-Brioux plain (350 km^2). It examines the contribution of grasslands to the conservation of various species of birds in a way compatible with an economically viable farming activity. Since the beginning of this research program, the surveyed zone has been progressively enlarged: 20 km^2 in 1994, 200 km^2 in 1995, 320 km^2 in 1996 with a relative stabilization until 2005 followed by other enlargements to 420 km^2 in 2006 and to 430 km^2 in 2007. In this paper, we consider the period of 12 successive years starting in 1996.

The long-term land-use survey

In the framework of its biodiversity research program in relation with agricultural practices, the Chizé Centre for Biological Studies (CEBC[2]) carries out, every year, two land-use surveys (in April and June). These two yearly surveys take into account both early harvested and late planted crops. In each survey, the occupancy of parcels is noted down distinguishing 47 land-uses (42 agricultural land-uses, 3 urban land-uses and 2 forest land-uses) as detailed in Table 1. Each parcel contains one type of crop and its physical boundaries can be a river, a path or a field limit. Each year, surveyors also update the parcel limits when a change is observed. The changing limits of the agricultural parcels led to the definition of elementary parcels, which resulted from a spatial union of previously updated parcel limits. The study area contains 19,000 elementary parcels, covering an area of 430 km^2. The collected data is stored in a GIS geodatabase, in vector format.

[1] Centre National de la Recherche Scientifique.

[2] Centre National de la Recherche Scientifique, Centre d'Etudes Biologiques de Chizé.

Table 1 Composition and average frequencies of adopted land-use categories

Land-use category	Land-use	Frequency (confidence, $P = 0.05$)	Cumulative frequency
Wheat (W)	Wheat, bearded wheat, cereal[a]	0.337 (±0.002)	0.337
Sunflower (S)	Sunflower, ryegrass followed by sunflower	0.139 (±0.001)	0.476
Rapeseed (R)	Rapeseed	0.124 (±0.001)	0.600
Urban (U)	Built area, peri-village, road	0.096 (±0.001)	0.696
Grassland and alfalfa (G)	Permanent grassland, grassland first year, temporary grassland (2–3 years), grassland of unknown age, alfalfa 1st year, alfalfa 2nd year, alfalfa 3rd year, alfalfa more than 3 years	0.078 (±0.001)	0.774
Maize (M)	Maize, rye grass followed by maize	0.076 (±0.001)	0.850
Forest and wasteland (F)	Forest or hedge, wasteland	0.034 (±0.001)	0.884
Winter barley (B)	Winter barley	0.034 (±0.001)	0.918
Ryegrass (Y)	Ryegrass, ryegrass followed by ryegrass	0.024 (±0.001)	0.942
Pea (P)	Pea	0.022 (±0.001)	0.964
Others (O)	Spring barley, grape vine, spontaneous fallow in June, foxtail millet, flax, oat, clover, field bean, rye grass followed by tillage, rye grass followed by unknown, spontaneous fallow followed by tillage, rye, cereal–legume mixture, spring crop, mustard, garden/market gardening, sorghum/millet, sorghum, millet, tillage, tobacco, other crop	0.036 (±0.001)	1.00

The "Frequency" refers to the average for the total area covered by all land-uses of a given land-use category. This average was computed over a 12-year period

[a] Cereal is used when the species can not be identified by the surveyor (it can be wheat, barley, ryegrass or other)

Theoretical background

Markov random fields and hidden markov models

Temporal and (or) spatial process can be spontaneously represented by stochastic graphs. The nodes are associated to some random variables (X, Y). The Xs take their values from a finite set of classes $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$—called the labels. The Ys take their values from the input data—also called the observations—of the process at a specific time slot or space location. The transitions represent the temporal or spatial dependencies between the nodes. The interest of such models is to compute the a posteriori probability

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$

for a given configuration of labels ($x_1, x_2, \ldots, x_n$) drawn from Ω and a set of observations ($y_1, y_2, \ldots, y_n$) in order to measure how the configuration fits the model assuming the observations. In a particular family of stochastic graphs—aka Markov Random Fields (MRF), the probability of observing $X_i = x_i$ assuming the values on all the other nodes depends on a limited set of nodes called the neighboring nodes of X_i.

$$P\left(X_i = x_i \middle| X_1 = x_1, \ldots, X_{j\neq i} = x_j, \ldots, X_n = x_n\right) = P\left(X_i = x_i \middle| \left\{X_{\text{neighbor of } i} = x_i\right\}\right)$$

In this paper, we present our work using second-order Hidden Markov Models (HMM2) to approximate MRF. We rely on the assumption that the distribution of land-use categories in an area at time t—the blocking plan—depends on the blocking plan observed at time $t - 1$ or $t - 2$ according to the order of the model (first-order or second-order). Hidden Markov Models analyze one dimensional sequences of observations. They differ from Markov chains (Castellazzi et al. 2008) through the presence of a supplementary hidden layer of nodes that models the structure of the data. This hidden layer is a second-order Markov chain that governs the sequence of random variables capturing the variability of the observations. Hidden Markov Models have been successfully introduced in 1976 in speech recognition (Jelinek 1976),

image processing (Benmiloud and Pieczynski 1995), biology (Hughey and Krogh 1996), and ecology (Le Ber et al. 2006). They can be used adequately to model temporal and spatial stochastic processes.

In order to model MRFs, Benmiloud and Pieczynski (1995) have proposed a method to convert a spatial representation of the data into a one dimensional sequence by means of a fractal curve—the Hilbert Peano curve—that spans the two dimensional space. Two points that are close to one another on the curve are close in the plane. The opposite is not true. Despite this drawback, they show that the classical HMM training algorithms give performances comparable to the more complicated algorithms involved in Markov Random Fields.

HMM2 definitions

An HMM2 is defined by:

- A set $S = \{s_1, s_2,\ldots,s_N\}$ of N states that are the outcomes of the variables X_t, where $t = 1,\ldots,T$.
- A transition matrix $A = (a_{ijk})$ over S^3, where a_{ijk} is the a priori transition probability $P(X_t = s_k \mid X_{t-2} = s_i, X_{t-1} = s_j)$ for the hidden Markov chain to be in state s_k at index t assuming it was in state s_j at index $t - 1$ and s_i at index $t - 2$. The Markov assumption states that these probabilities do not depend on index t.
- A set of N discrete distributions: $b_i(.)$ is the distribution of observations associated with state S_i. This distribution may be parametric, non parametric or even given by a HMM in a hierarchical HMM.

HMM2 properties

The second-order Hidden Markov Models are based on the probability and statistics theories. They implement an unsupervised training algorithm—the Baum–Welch algorithm—that estimates the HMM2 parameters from a corpus of observations. The estimated model enables to segment each sequence in stationary and transient parts and to build up a classification together with its a posteriori probability $P(X = \text{configuration} \mid Y = \text{observations})$.

In a first-order HMM, the probability that a sequence of n observations—called the state duration probability—is captured by state i follows a geometric decay defined by $(a_{ii})^n$, where a_{ii} is the a priori probability of the loop over state i. In a second-order HMM, the state duration is governed by two parameters, i.e., the probability of entering a state only once, and the probability of visiting a state at least twice, with the latter modeled as a geometric decay. This distribution better fits a probability density of durations than the exponential distribution of a first-order HMM. This property is of great interest when a HMM2 models a process in which a state captures only one or two observations (Mari et al. 1997).

Furthermore, the very success of HMMs is their robustness. Even when the input data do not fit a given HMM, it can give interesting results by discovering spatial and temporal regularities.

Hierarchical HMM

We model the spatial structure of the landscape by a MRF whose sites are random land-use sequences. These sequences are modeled by a temporal HMM2. This leads to the definition of a hierarchical HMM (HHMM) where a master HMM2 approximates the MRF. Then, the probability of a temporal land-use sequence is given by a temporal HMM2 as fully described by Fine et al. (1998) and Mari and Le Ber (2006). This hierarchical HMM is used to segment a landscape into regions. The temporal evolutions of the regional sites are land-use sequences that are modeled by a temporal HMM2. The use of hierarchical HMM in data mining is a special case of stochastic models as described by Le Ber et al. (2006) and may be summarized as the following steps:

1. Specify the topology of a hierarchical HMM;
2. Gather spatio-temporal data;
3. Train the HMM on these data using the Baum–Welch algorithm;
4. Segment the data and interpret the content of the classes;
5. Design a new model and go back to step 1.

ArpentAge: a data mining software

ARPENTAGE[3] (Analyse de Régularités dans les Paysages: Environnement, Territoires, Agronomie =

[3] http://www.loria.fr/~jfmari/App/

Analysis of Landscape Regularities: Environment, Territories and Agronomy) is an acronym that also means *landscape surveying* in French. It is the name of our knowledge discovery system based on higher-order hidden Markov models for analyzing spatio-temporal data bases. It takes as input an array of discrete data—the rows represent the spatial sites, the columns represent the time slots—and clusters the territory into patches whose crop sequences are extracted. This software allows the user to specify the architecture of the Markov model according to his objectives and the data. Displaying tools and the generation of shape files have also been implemented. The results of ArpentAge are interpreted and validated by domain specialists (i.e., agronomists).

A computer limitation issue when dealing with huge amounts of data

Using a Hilbert-Peano fractal curve requires regularly spaced input data points. This is why we rasterized land-use data by following a systematic and regular sampling pattern (10 m $\times$ 10 m). Data were then formatted so that the rows represent the spatial sites (sampled points) and the columns the time slots of attributes. A huge corpus with around 8 million rows has been obtained. However, the estimation of HMM2's parameters is a memory consuming process that can saturate even large computer memories. In order to help reduce the requirement of the memory resources, we choose to control two factors: (1) the length of the elementary Land-Use Successions (LUS), and (2) the size of the corpus of observations through the sampling resolution.

Results

Preliminaries: scaling the method to deal with huge amounts of data

Choice of the succession length

The length of Land-Use Successions (LUS) influences the interpretability of the final model. The longer the succession is, the more useful it is for agronomists. However, the total number of LUS is a power function of these succession lengths. Memory resources required during the estimation of HMM2 parameters increase with large numbers of LUS.

To help decide on the succession length, we compared the diversity of LUS between field-collected data and randomly generated data for different lengths of successions (Fig. 1). In the Niort Plain case, we can see that a 4-year succession length begins to differentiate from the random case. This reinforced our choice of the quadrennial succession as the elementary observation symbol in modelling the Niort Plain case study. Further down in this paper, the 4-year LUS will be sometimes referred to as quadruplets.

Choice of the spatial resolution

At regional scales, high-resolution samplings generate large amounts of data. With such amounts of input data, only rough models can be tested. On the other hand, with coarse resolution samplings, small parcels are omitted. In order to have an objective criterion for choosing the optimal spatial resolution, information loss in terms of LUS diversity was estimated for increasingly coarse resolution samplings. Figure 2 shows curves for the three considered LUS lengths that follow quite similar trends. As a compromise, we chose the (80 m $\times$ 80 m) resolution that led to a corpus 64 times smaller than the original one, with only a loss of 6% in information diversity.

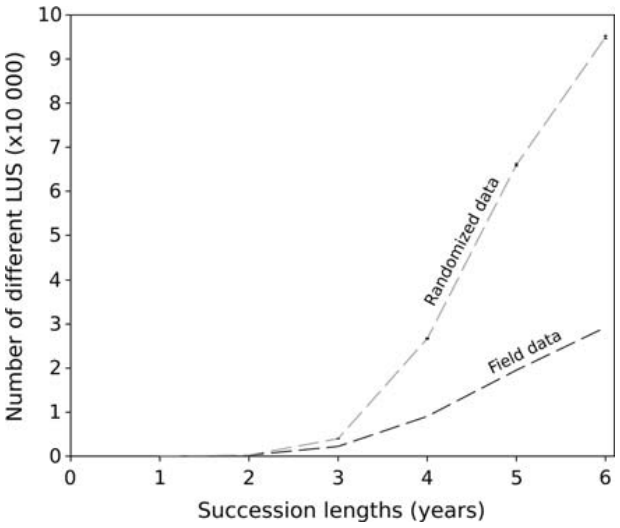


Fig. 1 Compared diversity of LUS between field-collected data and 10 random generated data sets for different succession lengths

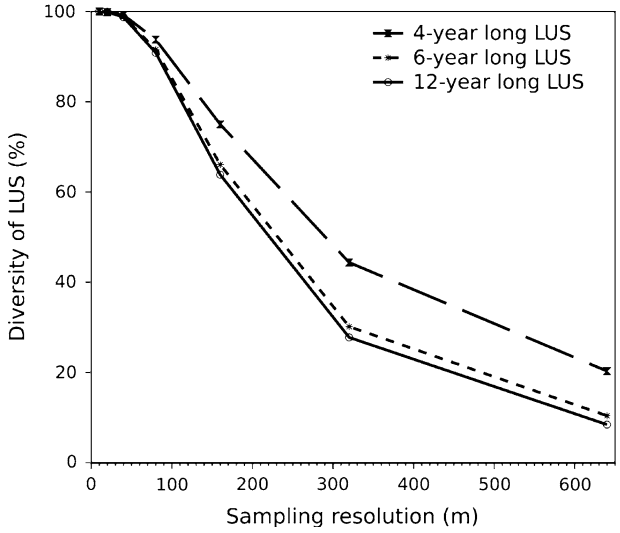


Fig. 2 Information loss in terms of LUS diversity in relation to sampling resolutions for three succession lengths. Tested resolutions are: 10, 20, 40, 80, 160, 320 and 640 m. Irregularity in sampling intervals is dictated by an algorithmic constraint, the resolution must be proportional to a power of 2. The most precise resolution is considered as the reference (100%)

Modelling the diversity of farming activity within landscapes

The approaches to model the diversity of farming activities differ according to whether one is considering the production system or the cropping system. Numerous classification models represent the diversity of agricultural production systems in a given region, the choice of which depends significantly on the time and space scales investigated. On the other hand, few models have been developed to represent the diversity of cropping systems on a regional scale. Here we propose an approach to model this diversity by first dealing with the Land-Use Successions (LUS) and then their locations.

The model representing the location of the diversity of LUS

The aim is to identify temporal stabilities in Land-Use Successions, and to locate them in landscapes. The first step was to build a typology of homogeneous land-use categories (Table 1). Then, we identified the successions in our landscape (Table 2). The third step identified the dynamics of LUS (Table 3; Fig. 3). The fourth step identified patches of LUS (Figs. 4, 5).

Table 2 Search pattern for extracting all 4-year long LUS involving one of the main land-use categories

1st Year	2nd Year	3rd Year	4th Year
X	?	?	?
?	X	?	?
?	?	X	?
?	?	?	X

X, the current main land-use category; ?, any land-use category

Land-use typology

In this first step we identified the major land-uses and classified them into homogeneous categories. Based on an arbitrarily frequency defined threshold (0.01 i.e., 1% of a given land-use among the total number of all land-use records in the data set), the land-use types were differentiated into major ($\geq$0.01) and minor ($<$0.01) land-uses. Then, major land-uses were grouped with other similar major or minor land-uses to form homogeneous land-use categories (hereafter called "main land-use categories"). Finally, the remaining minor land-uses were grouped into a residual category called "Others". This last land-use category is rather heterogeneous and will not be considered as a main land-use category in the following. Table 1 shows the result of this classification.

The conceptual model representing the diversity of LUS

Using 4-year long LUS to model the Niort Plain landscape required the reduction of the large number of these successions (many thousands, see Fig. 1). Most of the time, in the study period, the 10 main land-use categories almost describes the whole area of the Niort Plain (over 96%, see Table 1). On this basis, we chose to represent the diversity of LUS in as many classes as there are main land-use categories. Table 2 represents the search pattern used to extract all the successions involving a given main land-use category. For each main land-use category X, we look for the list S(X) of quadruplets in which X was involved.

Actually, Wheat (W), Sunflower (S) and Rapeseed (R) are generally integrated in a same succession (such as S–W–R–W, S–W–S–W, R–W–R–W, S–W–

Table 3 Results of the temporal analysis for the Niort Plain case study over the 1996–2007 period: frequency distribution of LUS for each class of successions

S(O)		S(W,S,R)		S(B)		S(U)		S(M)	
O–O–O–O	0.079	S–W–R–W	0.128	W–B–R–W	0.056	U–U–U–U	0.977	M–M–M–M	0.155
W–R–W–O	0.065	R–W–S–W	0.112	W–B–S–W	0.055	W–U–U–U	0.002	M–W–M–W	0.055
W–S–W–O	0.055	W–S–W–R	0.112	W–R–W–B	0.045	U–U–U–W	0.002	W–M–W–M	0.041
S–W–O–W	0.029	W–R–W–S	0.105	W–S–W–B	0.034	G–U–U–U	0.002	M–W–S–W	0.030
R–W–O–W	0.027	S–W–S–W	0.072	B–S–W–R	0.033	S–W–U–U	0.001	M–M–M–W	0.030
W–O–S–W	0.026	R–W–R–W	0.065	B–R–W–S	0.030	U–W–U–U	0.001	W–S–W–M	0.029
W–O–R–W	0.025	W–R–W–R	0.062	R–W–B–R	0.026	U–U–W–U	0.001	M–M–W–M	0.028
O–W–S–W	0.023	W–S–W–S	0.059	S–W–B–R	0.024	S–U–U–U	0.001	S–W–M–W	0.026
W–O–W–S	0.021	W–R–W–W	0.033	S–W–B–S	0.024	G–G–U–U	0.001	M–W–R–W	0.024
R–W–O–R	0.019	W–W–R–W	0.029	R–W–B–S	0.020	U–U–U–G	0.001	W–M–W–S	0.024

S(G)		S(F)		S(Y)		S(P)	
G–G–G–G	0.377	F–F–F–F	0.983	W–R–W–Y	0.045	P–W–R–W	0.052
G–G–G–W	0.040	F–F–F–W	0.002	W–S–W–Y	0.039	M–P–W–M	0.038
W–G–G–G	0.026	G–F–F–F	0.001	W–Y–S–W	0.038	W–P–W–R	0.033
Y–G–G–G	0.026	G–G–F–F	0.001	S–W–Y–W	0.028	P–W–S–W	0.031
S–G–G–G	0.025	F–F–F–S	0.001	Y–W–S–W	0.026	R–W–P–W	0.029
W–S–G–G	0.018	F–F–F–M	0.001	W–Y–W–S	0.025	W–S–W–P	0.027
G–G–G–S	0.016	F–F–M–M	0.001	R–W–Y–W	0.023	W–R–W–P	0.026
G–G–S–W	0.015	F–F–F–G	0.001	W–Y–W–R	0.019	M–M–P–W	0.026
O–G–G–G	0.015	Y–F–F–F	0.001	W–Y–W–Y	0.019	S–W–P–W	0.026
G–G–W–R	0.013	F–F–O–W	0.000	Y–W–R–W	0.018	P–W–M–M	0.022

Main land-use categories are: Wheat (W), Sunflower (S), Rapeseed (R), Urban (U), Maize (M), Grassland and alfalfa (G), Forest and wasteland (F), winter Barley (B), rYegrass (Y), Pea (P). O is a residual land-use category that contains successions of less frequent land-use types. Only the 10 most frequent successions are given

W–W, R–W–W–W, etc.). So, making a common class of these three crops allowed better results. To deal with this case, the above notation can be generalized to: S(*X*, *Y*,…) to denote the class of successions that involve at least one of the main land-use categories X, Y, etc.

We listed the resulting classes of successions as quadruplets "of interest" that we quantified by their frequency in the data corpus. The large number of quadruplets can then be reduced by using an appropriate threshold of cumulative frequency or by choosing a given number of most frequent quadruplets.

The LUS temporal dynamics analysis by means of HMM2

Table 3 shows the results of the temporal data mining analysis performed by the search patterns described in Table 2. Table 3 summarizes the distribution of quadruplets within the resulting classes of LUS.

Classes S(U) and S(F) are stable classes, they are mainly represented by only one quadruplet: U–U–U–U and F–F–F–F, respectively.

S(W,S,R) is the most dominant class. Its quadruplet distribution allows the deduction of principal rotations such as: (1) "SWRW" which is a quadrennial rotation composed of the four circular permutations of quadruplets: S–W–R–W, R–W–S–W, W–S–W–R, W–R–W–S. Its frequency is around 40%; and (2) "SW" and "RW": which are two biennial rotations deduced from the quadruplets: S–W–S–W, W–S–W–S and R–W–R–W, W–R–W–R, respectively. Their frequencies are slightly over 10%. These three rotations represent around 60% of the total composition of the S(W,S,R) class.

S(B), S(M), S(G), S(Y) and S(P) are less ordered classes and somehow reflect the farmers' "freedom"

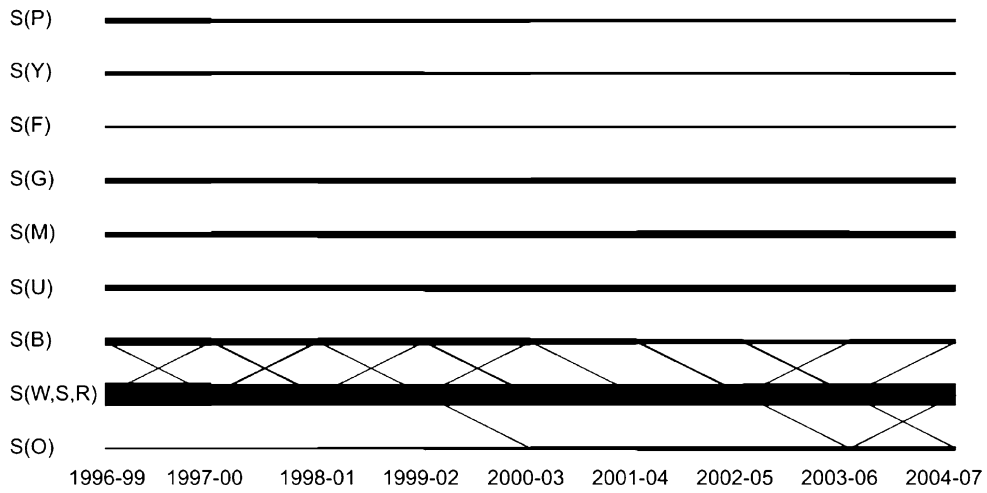


Fig. 3 Results of the temporal analysis for the Niort Plain case study over the 1996–2007 period: Markov diagram shows the temporal dynamics of LUS classes after 50 iterations of the EM algorithm. The *x*-axis represents the study period divided into sub periods of 4 years. The *y*-axis represents the classes of LUS involving: Wheat (*W*), Sunflower (*S*), Rapeseed (*R*), Urban (*U*), Maize (*M*), Grassland and alfalfa (*G*), Forest and wasteland (*F*), winter Barley (*B*), rYegrass (*Y*), and Pea (*P*). Others (*O*) is a residual land-use category that contains successions of less frequent land-use types. Diagonal transitions stand for inter-annual changes. *Horizontal* transitions indicate inter-annual stability. Only transitions whose frequencies are greater than 0.01 are displayed. The *line width* reflects the a posteriori probability of the transition assuming the observation of the 12-year LUS

as to their land allocation with the corresponding categories of crops (namely winter Barley (B), Maize (M), Grassland and alfalfa (G), rYegrass (Y) and Pea (P)). This corroborates our choice in considering the succession classes—S(*X*)—as the regularities to be located in the next step, rather than considering rotations.

In the **S(M)** class, the quadruplet distribution allows to deduce 3 main regularities: (1) 15% of the maize monoculture shown by the M–M–M–M quadruplet, and (2) about 10% of the biennial rotation "WM" deduced from the quadruplets M–W–M–W and W–M–W–M, and (3) about 0.8% of the quadrennial rotation "MWSW".

In the **S(P)** class, the quadruplet distribution shows the quadrennial rotations "PWRW" and "PWSW" whose respective frequencies are roughly 10%.

In the **S(B)** class, the quadruplets distribution shows two triennial rotations "RWB" and "SWB" whose respective frequencies are around 10%.

In the **S(O)** class, the O–O–O–O LUS characterizes the parcels covered by vineyards or market-garden crops that, mainly, keep the same land cover during all the study period. Other quadruplets incorporating a O appear with lower frequencies and reflect that this land cover can appear randomly in numerous parcels. The mean blocking plan of the O occupation is roughly 3% over the 12 years study period. Therefore, O occupations disappear in the amount of W, S, R in these parcels because the O occupations are not integrated in some specific rotations. This leads to a distribution's estimation almost equal to the S(W,S,R) distribution. We found this result by computing the divergence between these two distributions and found them close to each other.

Figure 3 shows the dynamics of the resulting classes of LUS in the Markov diagram. The quite constant width of horizontal lines of the Markov diagram indicates that during the whole study period (1996–2007), no major change has affected the dynamics of LUS. This figure shows diagonal transition lines between S(B) and S(W,S,R) classes and between S(O) and S(W,S,R). In the Niort area, Barley is cultivated in 3-year rotations ("RWB" or "SWB"). When these 3-year rotations are over, the farmers can start new rotations spanning 2 years ("RW", "SW") or 4 years ("SBRW"). The 3-year rotations are related to the horizontal lines in the S(B) class whereas transitions from a 3-year rotation to a 2- or 4-year rotation trigger a diagonal line to the S(W,S,R) class.

The diagonal transitions between S(O) and S(W,S,R) are explained by the closeness of the two distributions.

The LUS spatial analysis by means of HHMM2

In this step, temporal regularities (i.e., classes of successions) found in the temporal analysis step were localized within the Niort Plain landscape. As we observed that LUS was stationary over the 1996–2007 period, we used a simple temporal HMM2 to represent the states of the hierarchical HMM2. This temporal model does not segment the study period but rather considers it as a whole. This model has two states. One describes the distribution of the quadruplets of interest related to the class S(X), the other state captures the distribution of the quadruplets in the neighborhood. The Markov field introduces a blur in the patch's frontier and in the patch estimation because a site is classified not only based on its temporal characteristics (the quadruplet succession) but depends now on the classification of the neighboring sites. The map of the Niort Plain shown in Fig. 4, is the result of the temporo-spatial Markov modelling process that defines a classification based on the LUS that have been observed between 1996 and 2007. The landscape is seen as patches of LUS. For each class of patches, there is a corresponding distribution of LUS, which is summarized in a diagram format (Fig. 5).

The spatial partitioning with the HHMM2 fails to locate the S(O) class within the Niort Plain landscape. The spatial analysis located two small patches covering only 0.05% of the classified area, which correspond to six elementary parcels. For visibility and significance considerations, we did not represent the corresponding patches in the resulting map.

The class of patches labeled (a) represents the S(W,S,R) class of LUS in almost 70% of the total area. In the remaining area there are residual successions i.e., successions that contain neither Wheat nor Sunflower nor Rapeseed.

The class of patches labeled (b) represents successions in the Urban category. In this class of patches, the Urban category is present in LUS in more than 60% of the total area. The blur introduced by the Markov field can be seen in this patch. The temporal analysis exhibited a 0.975 probability for

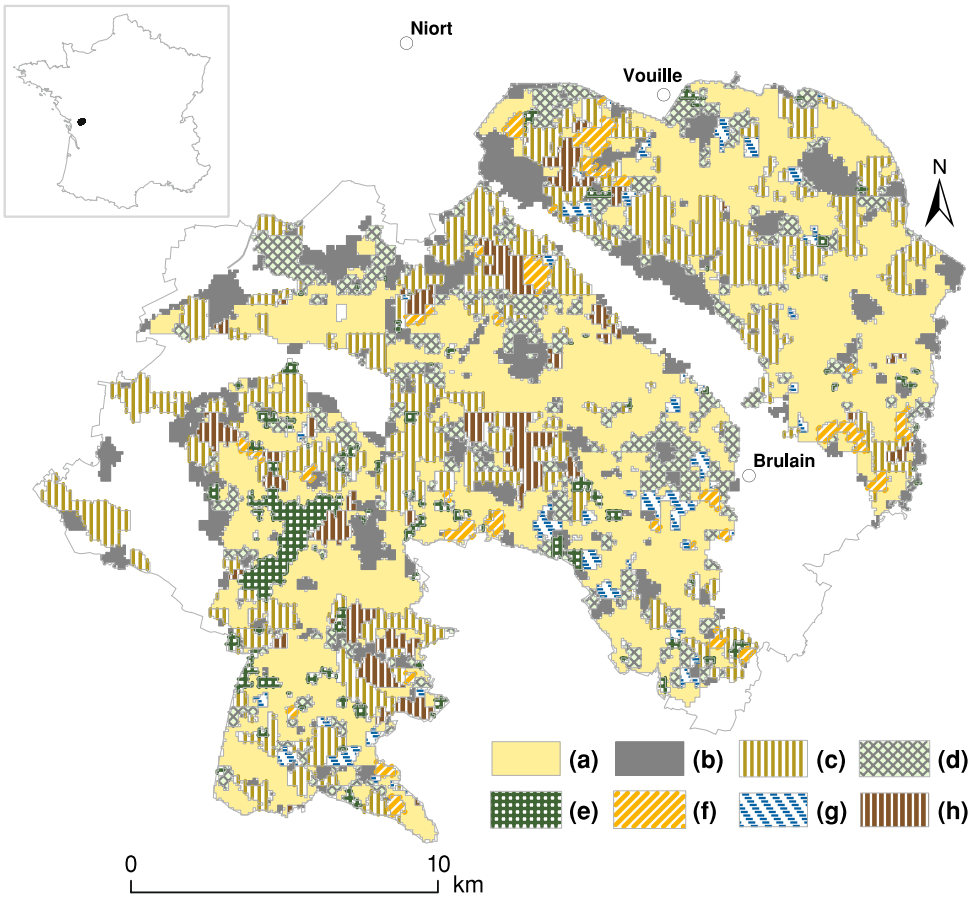


Fig. 4 The Niort Plain landscape as patches of LUS. *White areas* in the map are unclassified because they were insufficiently surveyed over the 1996–2007 period. Location of the Niort Plain in France is depicted in the *upper left-hand box*. The map legend is given in Fig. 5

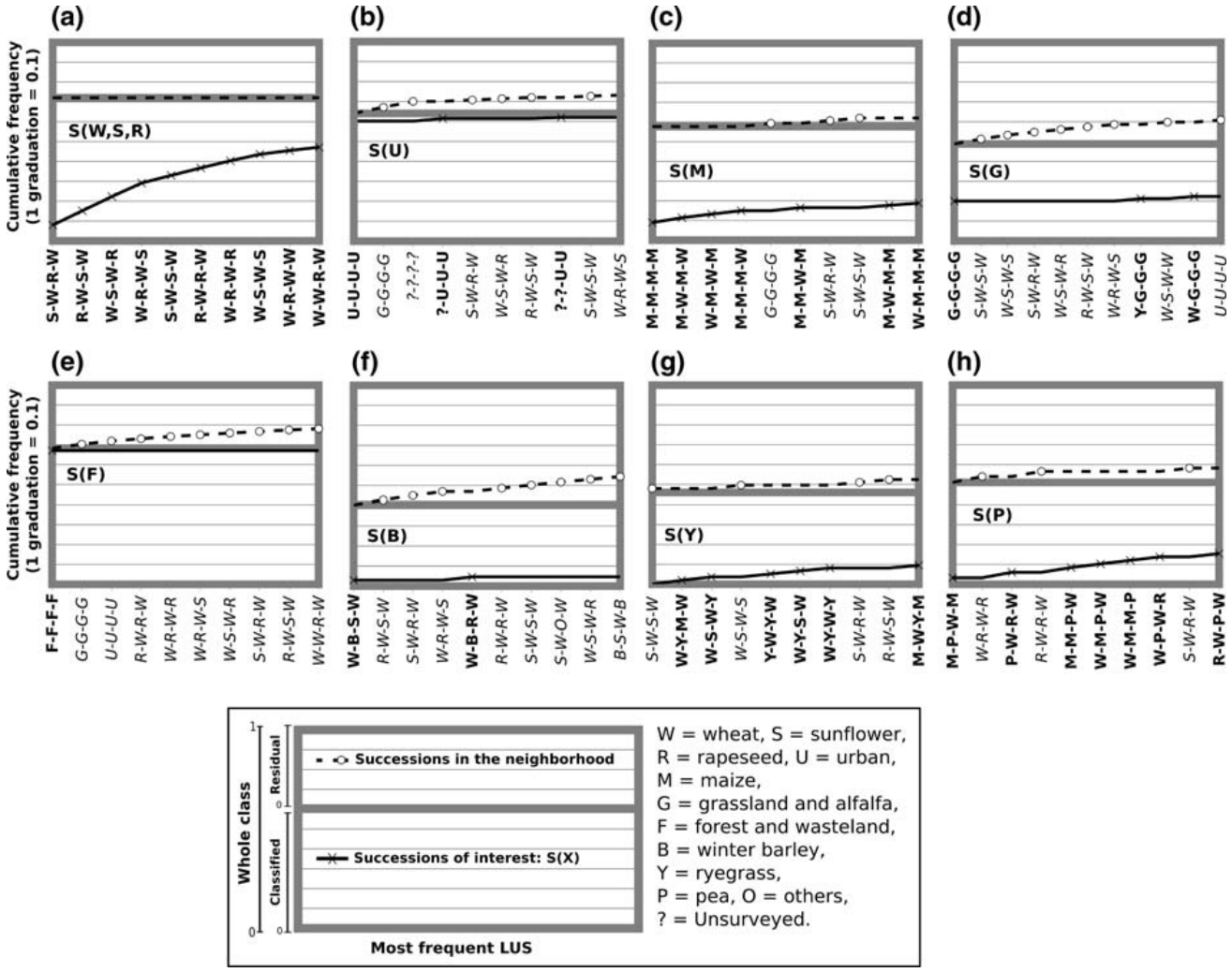


Fig. 5 Distribution of the most frequent LUS for each class of patches shown in Fig. 4. Each diagram plots separately the distribution of searched successions (*lower rectangle*) and residual successions (*upper rectangle*). The residual successions result from both *smoothing patches* and *border effects* introduced by the Markov field. *Bold* quadruplets belong to the successions of interest and *italic* quadruplets are residual successions. The frequency of each quadruplet has been computed by dividing the count of the quadruplet by the size of the class

the quadruplet U–U–U–U, the Markov field lowers this probability to 0.60. The residual part is mainly populated by the quadruplets of the S(G) class that are typical of urban neighborhoods. The same influence can be seen in the S(G) class where U–U–U–U is found in the residual part.

The map unit (c) represents the S(M) class of LUS in almost 60% of the total area, where around 10% has been cultivated with maize monoculture for at least 4 years (the quadruplet M–M–M–M).

The map unit (d) represents the S(G) class of LUS in almost 50% of the total area. It is mainly composed of old pastures (20% of G–G–G–G quadruplets), but also more recently converted to Grassland category areas (Y–G–G–G, W–G–G–G). Cropping systems using Wheat, Rapeseed and Sunflower are more likely to be found around grasslands.

The map unit (e) refers to Forest and Wasteland category in 70% of the total area of the associated patches. The only F–F–F–F quadruplet showing that Forests and Wastelands are rather stable represents this category. Close to forests, one is likely to find Grasslands, Urban areas and some cropping systems including Wheat, Rapeseed and Sunflower as shown in the residual part.

Map units (f,g,h) are less well classified. This is shown by their large residual intervals due to the neighborhood effects. However, the corresponding

Table 4 Density of Montagu's Harrier nests for each map unit

Map unit	Nests per km^2
a	3.42
b	0.00
c	0.97
d	0.51
e	0.00
f	1.07
g	2.79
h	1.38

The nests belong to the 1996–2007 period

patches are particularly interesting because they contain crops [respectively winter Barley (B), rYegrass (Y) and Pea (P)] whose frequency is more important than elsewhere.

Discussion

A new framework for landscape regularity modelling: a time × space analysis

The main stream of our research on landscape regularity modelling consists in a Time × Space analysis based on a stochastic approach. We pointed out the consistency of crop sequences (Le Ber et al. 2006; Castellazzi et al. 2008) by a LUS temporal analysis before locating the temporal regularities in the landscape by means of HHMM2.

The main advantages of this method are:

1. to be related with the farmers' choices since they use crop sequences instead of using a crop by crop organization,
2. to improve the landscape analysis with respect to field uses, and
3. to automatically learn the model parameters from the observation data.

In this paper, we described a new data mining method that processes a huge corpus of land-use observation data from a medium-size territory in order to:

1. choose the succession length of land-use,
2. choose the spatial resolution,
3. define a conceptual model for representing the diversity of LUS,
4. and finally, create maps that show patches of temporal regularities. These maps give an objective classification of the territory based on its pluriannual agronomic organization. They are of great interest when the ecological process under study and the spatial organization of the patches are correlated.

The need of perennial information systems

We used a data base built on a long term survey led by a CNRS team that involves a huge amount of human labor.

Another source of land-use maps is satellite remote sensing images, whose spatial and temporal resolutions have been greatly improved in recent years. As a common tool in Geography, it can be used to map land-uses on a regional scale (Girard et al. 1990; Veldkamp and Lambin 2001; Verburg and Veldkamp 2001). We put forward the hypothesis that our modelling method can handle those remote sensing data if they are able to inform a long time period with a sub-annual resolution.

Linking with biodiversity resources

To illustrate an attempt at linking agricultural patterns at a regional scale and ecological data, we chose a heritage species of birds: Montagu's Harrier. Nest locations of Montagu's Harrier overlapped the patches of LUS as depicted in Fig. 6.

The joint representation of nest locations and landscape organization is useful for ecologists to formulate research issues and hypotheses related to this bird's distribution over the agricultural area and over a longer period than 1 year. For instance, one can see that while patches of class (a) are homogeneously distributed over the studied area, one can wonder why the number of nests in the South-West patches of this class is so low compared to the mean density (Fig. 6 and Table 4). Ornithologists make the assumption that the presence of egg predators in the forest (patch e) prevent Montagu's Harriers from nesting in the vicinity and leads them to adopt a semi colonial behavior. But this statistical curiosity is still an open issue.

This illustration provides only broad tendencies on the joint relationship between Montagu's Harrier and its surrounding landscape. In fact, involved territories

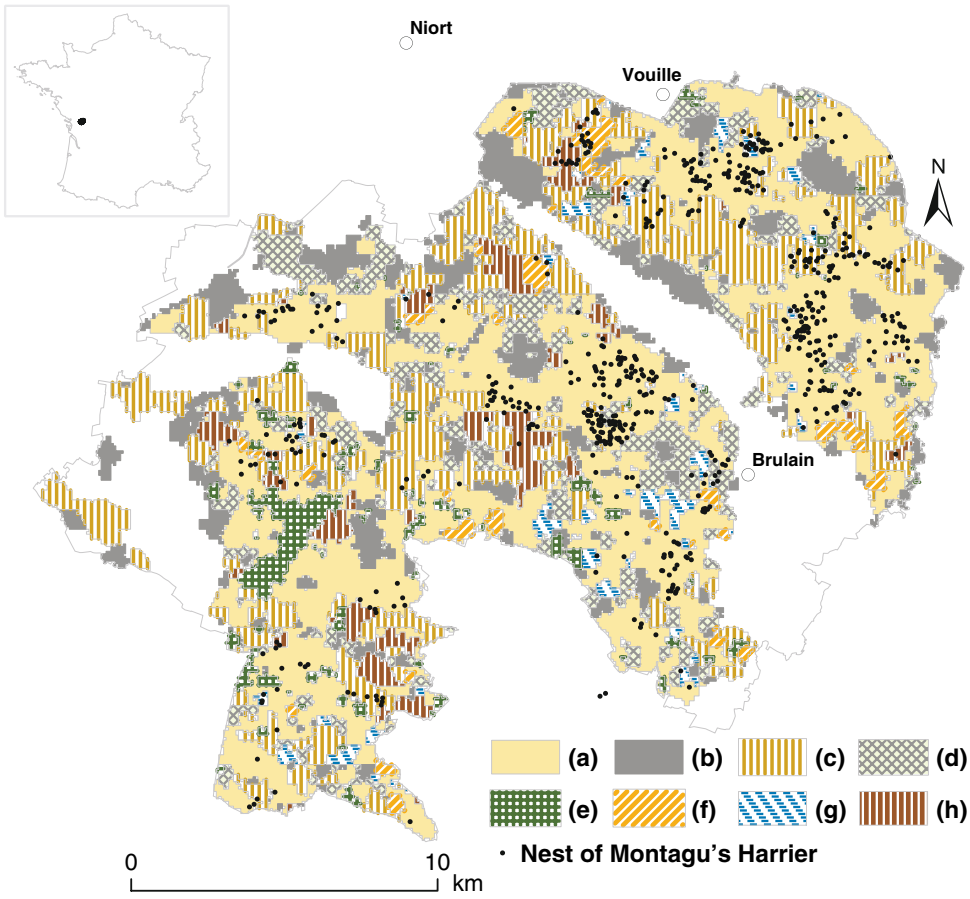


Fig. 6 Spatial distribution of Montagu's Harrier nests (*Circus pygargus*) within patches of LUS. The nests belong to the 1996–2007 period, which is the same period used to define patches of LUS. The location of nests results from a comprehensive monitoring throughout the study area each year from late April to late July

in the Montagu's Harrier's life are substantially larger than the nesting scale (Salamolard 1997). Landscape units have to be linked with animal territories and habitats. Future works will be to evaluate the areas involved in the definition of animal territories. Ecological and agricultural pattern interactions could be more deeply explored by following an organism-centered view of landscape heterogeneity (Turner 2005). This can be achieved by coupling agricultural and ecological data in the data mining process in order to extract joint regularities of both agricultural landscape descriptors and ecological indicators. Indeed, this needs a close collaboration with ecologists. This is what we intend to investigate in our future research work.

Contribution of agronomy to environmental issues at regional scales: contribution of landscape agronomy to landscape ecology

From an agronomist's point of view, we consider that agricultural landscapes are created by farmer practices (Morlon and Benoît 1990; Le Ber and Benoît 1998), and we seek to describe and to model the driving forces of landscape changes. In the present work, we hope to contribute to Landscape Agronomy which is an emerging field (Benoît et al. 2007) that holds three main scientific tasks:

1. to identify the rules and laws within the landscape that link environmental processes and farming systems;
2. to build scenarios for partners showing the implications of land-use practices for landscapes, and;
3. to build bridges between agronomists, geographers and ecologists on a common scientific field of interest: landscape.

As a contribution to Landscape Ecology, Landscape Agronomy is intended to give the "managing dimension" to the eco-field hypothesis of Farina and Belgrano (2006).

Acknowledgments This work was supported by the ANR-ADD-COPT project, the API-ECOGER project and the ANR-

BiodivAgrim project. We thank the CNRS team in Chizé for their data records obtained from their "Niort Plain data base". We thank Anne Mimet and the anonymous reviewers for their useful comments.

References

Benmiloud B, Pieczynski W (1995) Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images. Traitement du signal 12:433–454

Benoît M, Papy F (1997) Pratiques Agricoles et qualité de l'eau sur le territoire alimentant un captage. In: Neveu A, Riou C, Bonhomme R et al (eds) L'eau dans l'espace rural (Production végétale et qualité de l'eau). Aupelf-Uref-UREF & Inra éditions, Paris, pp 323–338

Benoît M, Mignolet C, Herrmann S et al (2007) Landscape as designed by farming systems: a challenge for landscape agronomists in Europe. In: Farming systems design 2007, methodologies for integrated analysis of farm production systems, Catania, 10–12 Sept 2007

Le Ber F, Benoît M (1998) Modelling the spatial organisation of land use in a farming territory. Example of a village in the Plateau lorrain. Agronomie 18:103–115

Butler SJ, Vickery JA, Norris K (2007) Farmland biodiversity and the footprint of agriculture. Science 315:381–384

Castellazzi M, Wood G, Burgess P et al (2008) A systematic representation of crop rotations. Agric Syst 97:26–33

De Koning GHJ, Verburg PH, Veldkamp A et al (1999) Multi-scale modelling of land use change dynamics in Ecuador. Agric Syst 61:77–93

Donald PF, Green RE, Heath MF (2001) Agricultural intensification and the collapse of Europe's farmland bird populations. Proc Roy Soc Lond 268:25–29

Farina A, Belgrano A (2006) The eco-field hypothesis: toward a cognitive landscape. Landscape Ecol 21:5–17

Fine S, Singer Y, Tishby N (1998) The hierarchical hidden markov model: analysis and applications. Mach Learn 32:41–62

Forman RTT, Godron M (1986) Landscape ecology. Wiley, New York

Fuller RM, Devereux BJ, Gillings S, Amable GS, Hill RA (2005) Indices of bird-habitat preference from field surveys of birds and remote sensing of land cover: a study of south-eastern England with wider implications for conservation and biodiversity assessment. Glob Ecol Biogeogr 14:223–240

Girard CM, Benoît M, De Vaubernier E et al (1990) SPOT HRV data to discriminate grassland quality. Int J Remote Sens 11:2253–2267

Glenn-Lewin DC, van der Maarel E (1992) Patterns and processes of vegetation dynamics. In: Peet RK, Glenn-Lewin DC, Veblen TT (eds) Plant succession: theory and prediction. Chapman & Hall, London, pp 11–44

Hughey R, Krogh A (1996) Hidden markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci 12:95–107

Jelinek F (1976) Continuous speech recognition by statistical methods. Proc IEEE 64:532–556

Landais E (1996) Typologies d'exploitations agricoles. Nouvelles questions, nouvelles méthodes. Economie Rurale (SFER) 236:3–15

Le Ber F, Benoît M, Schott C et al (2006) Studying crop sequences with CarrotAge, a HMM-based data mining software. Ecol Model 191:170–185

Leisz SJ, Thu Ha N, Bich Yen N et al (2005) Developing a methodology for identifying, mapping and potentially monitoring the distribution of general farming system types in Vietnam's northern mountain region. Agric Syst 85:340–363

Mari J-F, Le Ber F (2006) Temporal and spatial data mining with second-order hidden markov models. Soft Comput 10:406–414

Mari J-F, Haton J-P, Kriouile A (1997) Automatic word recognition based on second-order hidden markov models. IEEE Trans speech Audio Process 5:22–25

Mignolet C, Benoît M (2001) Réflexions sur une segmentation régionale selon la diversité des systèmes techniques agricoles—Cas de la plaine des Vosges. Géomatique 11:177–190

Mignolet C, Schott C, Benoît M (2007) Spatial dynamics of farming practices in the Seine basin: methods for agronomic approaches on a regional scale. Sci Total Environ 375:13–32

Moonen A-C, Bàrberi P (2008) Functional biodiversity: an agroecosystem approach. Agric Ecosyst Environ 127:7–21

Morlon P, Benoît M (1990) Étude méthodologique d'un parcellaire d'exploitation agricole en tant que système. Agronomie 6:499–508

Peet RK, Glenn-Lewin DC, Veblen TT (1992) Plant succession: theory and prediction. Chapman & Hall, London

Perrot C (1990) Typologie d'exploitations construite par agrégation autour de pôles définis à dire d'experts. Proposition méthodologique et premiers résultats obtenus en Haute-Marne. Prod Anim 3:51–66

Poudevigne I, Alard D (1997) Landscape and agricultural patterns in rural areas: a case study in the Brionne Basin, Normandy, France. J Environ Manag 50:335–349

Reboul C (1976) Mode de production et systèmes de culture et d'élevage. Economie Rurale 112:55–65

Retho B, Gaucherel C, Inchausti P et al (2008) Spatially explicit population dynamics of Pterostichus Melanarius I11. (Coleoptera: Carabidae) in response to changes in the composition and configuration of agricultural landscapes. Landsc Urban Plan 84:191–199

Robinson RA, Sutherland WJ (2002) Post-war changes in arable farming and biodiversity in Great Britain. J Appl Ecol 39:157–176

Salamolard M (1997) Utilisation de l'espace par le Busard cendré *Circus pygargus*. Superficie et distribution des zones de chasse. Alauda 65:307–320

Sebillotte M (1974) Agronomie et agriculture. Essai d'analyse des tâches de l'agronome. Cahiers de l'ORSTOM 24:3–25

Turner MG (1990) Spatial and temporal analysis of landscape pattern. Landscape Ecol 4:21–30

Turner MG (2005) Landscape ecology: what is the state of the science? Annu Rev Ecol Evol Syst 36:319–344

Usher MB (1992) Statistical models of succession. In: Peet RK, Glenn-Lewin DC, Veblen TT (eds) Plant succession:

theory and prediction. Chapman & Hall, London, pp 215–246

Veldkamp A, Fresco LO (1997) Reconstructing land use drivers and their spatial scale dependence for Costa Rica (1973 and 1984). Agric Syst 55:19–43

Veldkamp A, Lambin E (2001) Predicting land-use change. Agric Ecosyst Environ 85:1–6

Verburg PH, Veldkamp A (2001) The role of spatially explicit models in land-use change research: a case study for cropping patterns in China. Agric Ecosyst Environ 85:177–190

Welch LR (2003) Hidden markov models and the Baum-welch algorithm. IEEE Inform Theory Soc Newsl 53:10–13

Fouille de données à l'aide de modèles stochastiques

Segmentation temporo-spatiale des successions de cultures d'un territoire agricole à l'aide de HMM2

El Ghali Lazrak* — **Marc Benoît*** — **Jean-François Mari*,****

* *INRA, UR 055, SAD ASTER*
domaine du Joly, F-88500 Mirecourt
{lazrak, benoit}@mirecourt.inra.fr

** *UMR CNRS 7503 et INRIA-Grand Est, LORIA*
B.P. 239 F-54506 Vandœuvre-lès-Nancy
jfmari@loria.fr

RÉSUMÉ. *Dans un contexte de développement durable, l'activité agricole doit préserver le milieu naturel qu'elle utilise et transforme. Représenter la répartition de l'activité agricole dans l'espace et le temps est une des premières étapes pour préserver et rétablir l'équilibre dans les agro-écosystèmes. Nous modélisons l'activité agricole d'un méso agro-écosystème en utilisant la première composante du système de culture, à savoir les successions de cultures. Nous considérons le territoire agricole comme une mosaïque de parcelles sur lesquelles se trouvent des couverts végétaux. La méthode consiste à : (i) choisir la longueur de la succession temporelle du sol (OCS) ; (ii) choisir la résolution de l'échantillonnage spatial ; (iii) définir un modèle conceptuel pour la représentation des successions des OCS ; (iv) créer une carte de paysages sous forme de zones homogènes vis-à-vis des successions d'occupation du sol et croiser enfin cette information avec la présence / absence des espèces animales considérées ce qui dégage des zones à prospecter plus profondément.*

ABSTRACT. *We propose an original data mining method to segment agricultural landscapes based on a temporo spatial modelling of their crop successions. This method consists in (i) choosing the length of the crop succession, (ii) choosing the spatial resolution to sample the territory, (iii) specifying a conceptual model to represent the crop successions by means of second-order Hidden Markov Models, and (iv) finally segmenting the territory into homogeneous areas that will be further investigated.*

MOTS-CLÉS : `HMM2`*, segmentation temporo-spatiale, fouille de données*

KEYWORDS: `HMM2`*, temporo-spatial segmentation, data mining*

1. Introduction

Dans un contexte de développement durable, l'activité agricole se doit, entre autres exigences, de préserver le milieu naturel qu'elle utilise et transforme. Représenter la répartition de l'activité agricole dans l'espace à des échelles compatibles avec celles du déroulement d'enjeux environnementaux et/ou de processus écologiques va dans le sens des efforts visant à préserver et à rétablir l'équilibre dans les agro-écosystèmes [P.T 99].

L'étude des successions de couverts végétaux pour représenter l'activité agricole d'un agro-système a été initiée depuis la fin du XIXe siècle dans le cadre de travaux sur les dynamiques de la végétation dans des écosystèmes naturels [GLE 92]. Les successions de cultures intègrent différentes logiques qui dictent le travail de l'agriculteur. La fouille des successions de cultures d'un territoire a pour objectif de mettre à jour ces logiques et de quantifier leur importance sur l'évolution des paysages et l'impact des contraintes agricoles sur la biodiversité.

La méthode de fouille de données présentée dans cet article considère le territoire agricole comme une mosaïque de parcelles sur lesquelles se trouvent des couverts végétaux ou d'autres occupations telles que bâti, routes, …. La méthode consiste à : (i) choisir la longueur de la succession temporelle du sol (OCS) ; (ii) choisir la résolution de l'échantillonnage spatial ; (iii) définir un modèle conceptuel pour la représentation des successions des OCS ; (iv) créer une carte de paysages sous forme de zones homogènes vis-à-vis des successions d'occupation du sol et croiser enfin cette information avec la présence / absence des espèces animales considérées ce qui dégage des zones à prospecter plus profondément.

Après l'introduction suivie de la présentation du matériel et des méthodes, la partie 3 décrit notre application dans laquelle nous recherchons le lien entre successions de cultures et présence d'une espèce animale protégée en l'occurrence le busard de Montagu, dans un territoire de 350 Km^2 dans l'Ouest de la France. La conclusion esquisse la suite de ce travail préliminaire de fouille de données temporelles et spatiales.

2. Matériels et méthodes

2.1. *Les occupations du sol*

Le territoire agricole étudié – 350 Km^2 dans la plaine céréalière de Niort – est enquêté depuis plus de 12 ans. La localisation et les occupations de ses parcelles sont relevées chaque année. Cette enquête est stockée dans un système d'information géographique (SIG) et est destinée à suivre les évolutions des occupations et des rotations de cultures notamment en ce qui concerne l'évolution des prairies : OCS essentielles pour la préservation de certaines espèces animales protégées. L'ensemble constitue un gros corpus d'informations temporelles et spatiales possédant un niveau de détails supérieur à ce qu'une analyse d'images satellites peut actuellement obtenir. Ce corpus

est représenté sous la forme d'une matrice dans laquelle les colonnes représentent les OCS année par année et les lignes représentent les différents sites enquêtés.

L'analyse des fréquences moyennes annuelles des OCS calculées sur les 12 années de la période d'étude fait ressortir 47 occupations du sol de la matrice de données. Les experts agronomes les regroupent ensuite en 10 catégories (*cf.* tableau 1) suivant une démarche tenant compte de la similitude des conduites culturales. Sont retenues les OCS : Blé (B), Tournesol (T), Colza (C), Urbain (U), Prairies et luzernes (P), Maïs (M), Forêts et friches (F), Orge d'hiver (O), raY-grass (Y), poiS (S) et Autres (A).

Catégorie d'occupation du sol	Occupation du sol	Fréquence cumulée
Blé (B)	blé, blé barbu, céréale	0.337
Tournesol (T)	tournesol, ray-grass suivi de tournesol	0.476
Colza (C)	Colza	0.600
Urbain (U)	bâti, péri-village, route	0.696
Prairies et Luzernes (P)	prairie permanente, prairie année 1, prairie temporaire (2-3 ans), prairie âge inconnu, luzerne 1 an, luzerne 2 ans, luzerne 3 ans, luzerne > 3 ans	0.6
Maïs (M)	maïs, ray-grass suivi de maïs	0.850
Forêts et friches (F)	forêt ou haie, friche	0.884
Orge d'hiver (O)	orge d'hiver	0.918
Ray-grass (Y)	ray-grass, ray-grass suivi de ray-grass	0.942
Pois (S)	Pois	0.964
Autres (A)	orge de printemps, vigne, jachère spontanée (juin), moha, lin, avoine, trèfle, féverole, ray-grass suivi de labour, ray-grass suivi d'inconnu, jachère spontanée suivie de labour, mélange céréale légumineuse, culture printemps, moutarde, jardin / culture maraîchère, sorgho / millet, sorgho, millet, labour, tabac, autre culture	1.000

Tableau 1: Composition et fréquences moyennes des catégories d'occupation du sol adoptées

2.2. *La mosaïque agricole*

Le composant de base dans un territoire agricole est la parcelle : polygone de taille variable possédant un couvert – l'OCS – et éventuellement habitée par une espèce animale. Elle est délimitée géographiquement par une route, un chemin, … ou artificiellement par l'agriculteur qui subdivise le territoire de son exploitation pour respecter un assolement : ensemble des surfaces allouées à chaque culture. Les fron-

tières des parcelles changent chaque année en fonction des choix des agriculteurs (*cf.* figure 1). Pour tenir compte de ce changement, les enquêteurs définissent l'ensemble des micro-parcelles comme étant l'union de toutes les intersections de parcelles pendant la période d'étude. Il y a environ 20000 micro-parcelles dans le territoire étudié. Tous les points d'une micro-parcelle n'ont hébergé qu'une succession de cultures pendant la période d'étude.

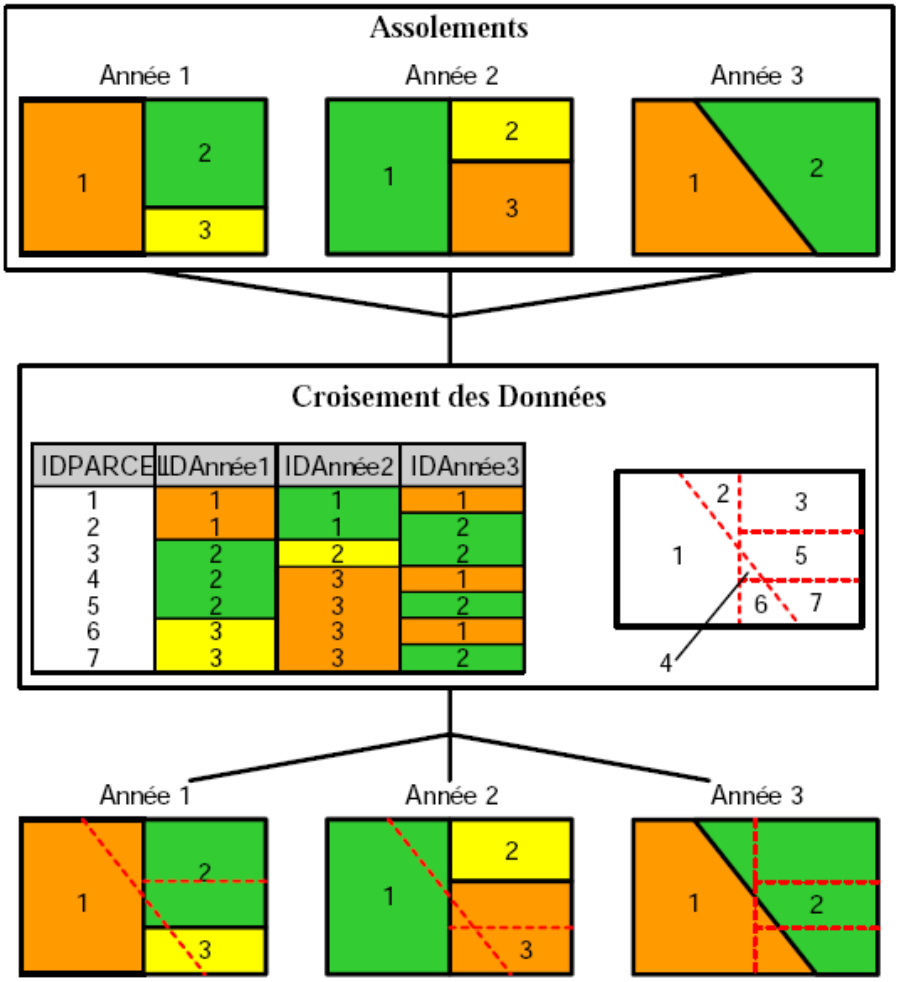


Figure 1. *Exemple d'évolution des limites de parcelles pendant trois années successives. L'intersection des parcelles pendant cette période aboutit à la définition de sept micro-parcelles*

L'analyse spatiale d'un territoire agricole vu comme une mosaïque de parcelles de taille quelconque présente des particularités dues aux caractéristiques des OCS. Dans l'analyse spatiale du territoire, la parcelle joue le rôle d'un pixel de taille variable possédant plusieurs modalités : les différentes OCS. L'OCS d'une parcelle à l'instant t s'inscrit dans une succession temporelle d'OCS qui intègre le savoir-faire de l'agriculteur qui met en valeur son territoire en fonction de différentes logiques. L'organisation dans le temps implique une organisation dans l'espace. Ainsi l'OCS d'une parcelle une année donnée dépend de l'OCS les années précédentes sur cette parcelle ainsi que des OCS des parcelles voisines. Dans la mosaïque parcellaire, le système de voisinage est irrégulier. Une parcelle a un nombre quelconque de parcelles avec lesquelles elle partage une frontière commune. Nous modélisons la mosaïque parcellaire par un champ de Markov irrégulier dans lequel la parcelle s'inscrit dans une succession temporelle d'OCS. L'ensemble constitue une image de successions plutôt qu'une succession d'images et ne peut être étudié directement par les modèles numériques d'images appliqués au suivi de trajectoires d'objets [NOY 07]. Enfin, les territoires étudiés ne sont pas carrés et les parcelles ajoutées pour leur donner une forme carrée doivent être en nombre minimum afin de ne pas perturber la classification spatiale.

2.3. *Modèles stochastiques pour représenter les successions de cultures de la mosaïque parcellaire*

2.3.1. *Modèles pour représenter l'évolution temporelle*

Les successions de couverts végétaux se prêtent bien à la modélisation avec des modèles de Markov [LEB 06, CAS 08]. Plus l'ordre du modèle de Markov est long, plus celui ci peut modéliser de longues successions et plus il intéresse l'agronome. Les paramètres du modèle de Markov peuvent être estimés par expertise [CAS 08], comme ils peuvent être automatiquement calculés à partir d'un corpus d'observations en utilisant des algorithmes d'apprentissage comme celui de Baum-Welch [BAU 72] lorsqu'il s'agit de Modèles de Markov cachés (HMM). Dans notre cas, nous avons reproduit en l'adaptant la démarche de fouille effectuée sur de grands territoires en utilisant des HMM d'ordre 2 (`HMM2`) capables de modéliser [MAR 02, BEN 03, LEB 06] des rotations de cultures sur plus de deux ans, ce qui constitue un avantage sur les HMM d'ordre 1. Le HMM a l'avantage – comparé à la chaîne de Markov – de représenter l'état du processus par une distribution sur l'ensemble des observations plutôt que par une seule observation comme c'est le cas dans une chaîne de Markov. L'observation est habituellement une OCS ou une succession temporelle de 2 à 4 OCS. Ceci permet plus de souplesse. Enfin les HMM permettent une modélisation spatiale. Ce travail montre nos résultats en classification en utilisant des `HMM2` aussi bien pour traiter la dimension temporelle que la dimension spatiale des données, ce qui donne plus de cohérence aux traitements.

6 Colloque STIC et Environnement.

2.3.2. *Modèles pour représenter un champ spatial de Markov*

Les champs de Markov permettent une classification non supervisée d'images numériques [JUL 75, GEM 84, BAR 00]. Pour contrer la puissance de calcul nécessaire à leurs estimations, des méthodes approchées à l'aide de courbes fractales parcourant le plan et segmentées par un HMM ont été développées [BEN 95, DER 04]. Le modèle utilisé dans cet article pour segmenter l'espace utilise la dernière approche et s'appuie sur un HMM2 proposé [MAR 02, LEB 06] dans le cadre de l'analyse des données spatio-temporelles *Ter-Uti* qui définissent un maillage spatial régulier de la France. Nous l'avons adapté pour traiter des pixels de taille différente - les micro-parcelles – dans un système de voisinage irrégulier. Nous utilisons une courbe fractale de Peano dont la "profondeur de fractalisation" s'ajuste à la taille de la micro-parcelle. Un processus récursif fusionne 4 points situés sur la courbe quand ils correspondent au motif générateur et qu'ils sont dans la même micro-parcelle. La figure 2 donne un exemple du parcours de la fractale dans une mosaïque parcellaire échantillonnée par une image 16 x 16. Dans la suite des traitements spatiaux, le territoire sera représenté par la suite des points de la fractale.

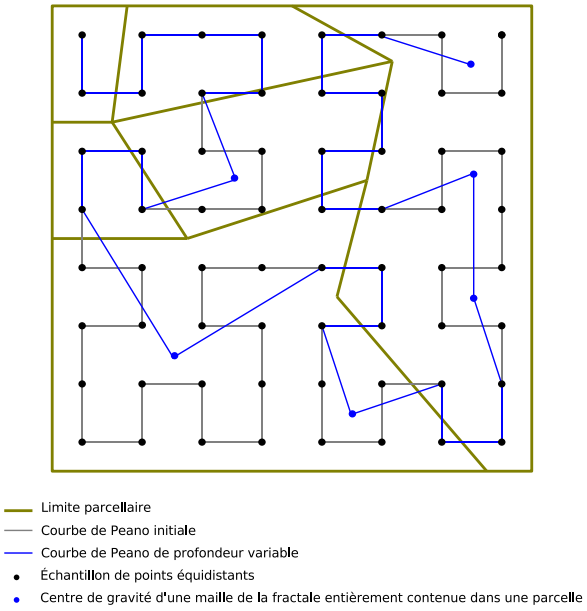


Figure 2. *Prise en compte de la taille des parcelles dans le parcours du plan par la courbe fractale de Hilbert-Peano. L'exemple montre deux fusions successives dans la parcelle située en bas à gauche aboutissant à l'agglomération de 16 points en un seul*

2.3.3. *Modèles pour représenter un champ spatial de successions : les HMM hiérarchiques*

Les HMM2 temporels et HMM2 spatiaux sont fusionnés dans un modèle de Markov caché hiérarchique (HHMM 2) [JEN 01, MAR 07] et segmentent la représentation spatiale du territoire en zones homogènes, c'est à dire constituées de parcelles dont la succession d'OCS est modélisée par le même HMM2. La figure 3 donne un exemple d'HMM2. L'ensemble des programmes permettant la spécification des HHMM2, leur apprentissage à l'aide d'un corpus de données, la segmentation en zones homogènes représentées par les états cachés de HMM2 ainsi que la création des shapefiles visualisant ces zones constituent la boîte à outils ARPENTAGE (Analyse de Régularités dans les Paysages : Environnement, Territoires et Agronomie), sur ensemble de CARROTAGE [1] utilisé pour l'analyse des successions d'OCS.

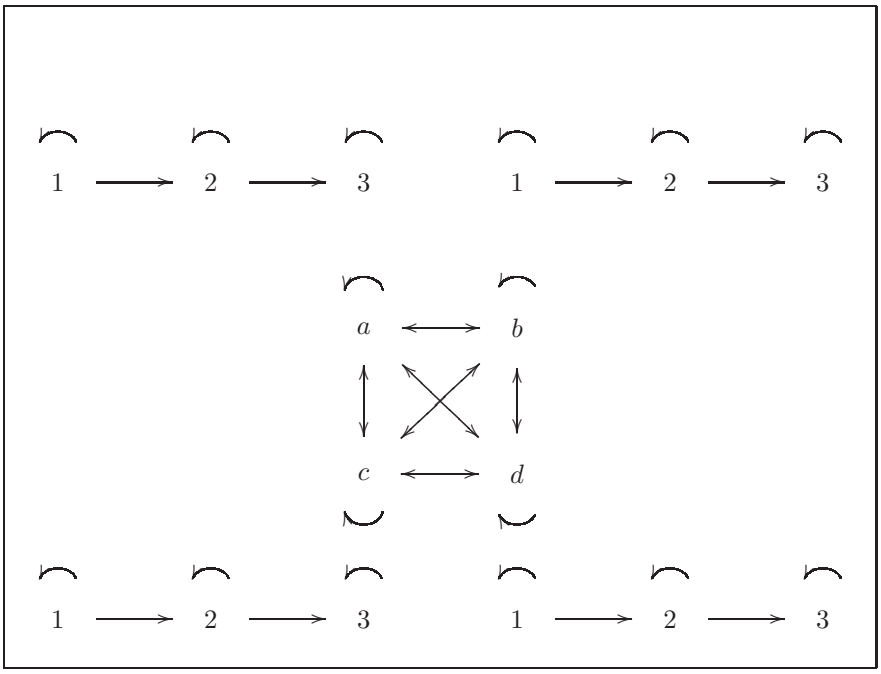


Figure 3. *Chaque état a, b, c, d du HHMM2 est un HMM2 temporel dont les états sont 1, 2, 3.*

3. Application

Notre but est d'identifier les régularités temporelles en termes de successions d'OCS puis de les localiser dans le territoire afin de croiser les zones homogènes

1. licence GPL

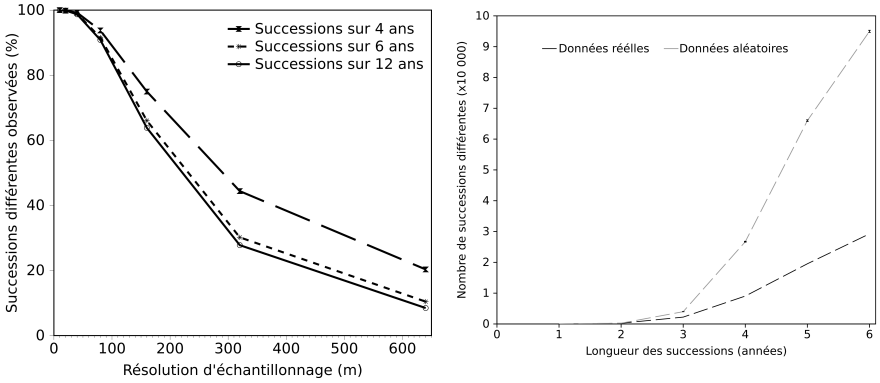
découvertes (“patches”) avec la localisation des nids de busards de Montagu – espèce animale protégée – connue également des enquêteurs.

3.1. *Choix de la longueur des successions et de la résolution d'échantillonnage*

La charge de calcul de l'estimation des paramètres des `HMM` peut être réduite en contrôlant deux facteurs : (i) la taille de la matrice des données à travers la résolution d'échantillonnage spatial, (ii) et la longueur de la succession d'occupation du sol.

Afin de disposer d'un critère objectif pour le choix de la résolution spatiale, la perte d'information en terme de diversité des successions d'occupation du sol a été quantifiée pour des résolutions d'échantillonnage de moins en moins élevées (*cf.* figure 4(a)). On remarque qu'une résolution de 80 m x 80 m permet d'obtenir une matrice de données 64 fois plus réduite que la matrice de données initiale avec une perte de seulement 6% en termes de diversité d'information.

Nous nous sommes intéressés ensuite à l'influence de la longueur des successions sur l'entropie du système. Nous avons comparé le nombre de successions différentes à celles produites aléatoirement dans le même territoire. Chaque OCS est tirée selon une loi uniforme sans tenir compte ni de sa localisation ni des OCS précédentes au même endroit. La figure 4(b) montre qu'à partir d'une longueur de succession de 4 années, la zone d'étude commence à se distinguer nettement du modèle où les successions sont choisies de manière aléatoire dans la zone d'étude.



(a) Nbr. de successions vs résolution spatiale (b) Nbr. de successions vs durée de la succession

Figure 4. *(a) Évolution de la diversité spatiale en fonction de la résolution d'échantillonnage. (b) Comparaison de la diversité des successions d'occupation du sol entre les données réelles d'occupation du sol et des données générées aléatoirement pour différentes longueurs de successions*

3.2. *Influence de la résolution spatiale sur la longueur de la courbe fractale*

Le SIG permet de définir une image raster en appliquant une grille de points régulièrement espacés. Le parcours de Hilbert-Peano sur une grille de points se détermine facilement quand la grille est un carré de coté 2^n. Différentes tailles d'image – de 4096 à 256 correspondant à des résolutions spatiales de 10 à 160 mètres – ont été choisies pour contenir le territoire d'étude. Pour insérer le territoire d'étude dans une image carrée, nous l'entourons d'une "grande" micro-parcelle constituée de points "blancs" possédant l'occupation "Hors zone". La table 2 montre la réduction de la longueur de la fractale en fonction de la résolution spatiale. La colonne 4 donne le nombre de points de l'image situés sur le territoire.

long. fractale	taille image	résolution (m)	nb. points échantillonnés
1160245	4096	10	4287149
481462	2048	20	1074298
182476	1024	40	269832
61285	512	80	68052
17830	256	160	17278

Tableau 2. *Longueur de la fractale en fonction de la résolution spatiale adoptée pour représenter le territoire*

On remarque qu'une résolution de 160 mètres ne procure aucun gain ; il n'y pas eu d'agglomération de points à l'intérieur d'une même micro-parcelle sauf pour la micro-parcelle "blanche" constituée des points "Hors zone". La fractale passe par tous les points échantillonnés du territoire. Environ 95% des micro-parcelles ont moins de 320 mètres de coté et donc échappent à tout regroupement de points. A partir d'une résolution de 40 mètres (70% des micro-parcelles ont plus de 80 mètres de coté) l'intérêt du regroupement des points apparaît nettement. Un inconvénient de la fractalisation à profondeur variable est que le nombre de points ne rend plus compte facilement de la surface. Les distributions d'OCS dans une zone donnée ne sont plus des assolements.

3.3. *Recherche des régularités temporelles*

Nous représentons la diversité des successions d'occupation du sol en autant de classes qu'il y a de catégories d'occupation du sol. Les classes de successions sont notées S(X), X étant une catégorie d'OCS, et la classe S(X) se lit : classe des successions avec X. Le tableau 4 représente le modèle de recherche utilisé pour extraire les successions appartenant à une classe d'occupation donnée. Une classe d'occupation se présente donc sous forme d'une liste de quadruplets associés chacun à sa fréquence dans la matrice de données. Chaque quadruplet comporte, au moins dans un de ses 4 éléments constitutifs, une occurrence de la catégorie d'occupation du sol ayant donné son nom à la classe.

Dans le territoire d'étude, le Blé (B), le Tournesol (T) et le Colza (C) sont le plus souvent intégrés dans une même succession de 4 ans (p. ex. TBCB, TBTB, CBCB, TBBB, CBBB). Les fouilles de données avec une classe commune à ces trois cultures ont permis d'obtenir des résultats plus cohérents qu'avec les classes séparées. Cette classe est notée S(B,T,C) regroupe les quadruplets impliquant au moins l'une de ces trois cultures.

Année t	Année t+1	Année t+2	Année t+3
X	?	?	?
?	X	?	?
?	?	X	?
?	?	?	X

Tableau 4. *Le motif de recherche pour l'extraction des quadruplets impliquant une catégorie X d'occupation du sol. ? représente une quelconque catégorie d'occupation du sol*

3.4. *Analyse temporelle*

Suite aux résultats trouvés dans le paragraphe 3.1, nous avons utilisé une résolution spatiale de 80 mètres et des observations constituées de 4 OCS temporelles successives sur tous les points du territoire. Les régularités temporelles ont été recherchées par fouille de données à l'aide de `HMM2` dont les états ont été initialisés à l'aide de distributions de quadruplets d'OCS correspondant aux motifs de recherche décrits dans le tableau 4. Le `HMM2` possède 10 états : autant que de S(X), X étant une des 10 catégories d'OCS (*cf.* tableau 1). Tous les états sont inter connectés. Ce type de modèle correspond au modèle de "type 2" décrit dans [LEB 06]. Un apprentissage de ce `HMM2` par l'algorithme de Baum-Welch détermine les probabilités a posteriori de transitions entre états comme l'illustre la figure 5. Cette figure montre une monotonie paysagère en terme de successions d'occupation du sol. En effet, la largeur quasi constante des lignes horizontales et diagonales du diagramme indique que durant la période d'étude (1996-2007), aucun changement majeur n'a affecté la dynamique des successions d'occupation des sols dans la zone d'étude. Ceci nous a conduit, dans l'étape de spatialisation des régularités temporelles à considérer stationnaire le processus d'allocation des terres en terme de succession d'occupation du sol.

3.5. *Résultats de la classification temporo spatiale*

Dans cette étape, nous localisons les régularités temporelles trouvées dans l'analyse temporelle. Comme le territoire ne présente aucune dynamique temporelle (*cf.* figure 5), chaque `HMM2` temporel du `HHMM2` ne contient qu'un seul état initialisé par les distributions trouvées dans l'analyse temporelle. La segmentation spatiale est faite par

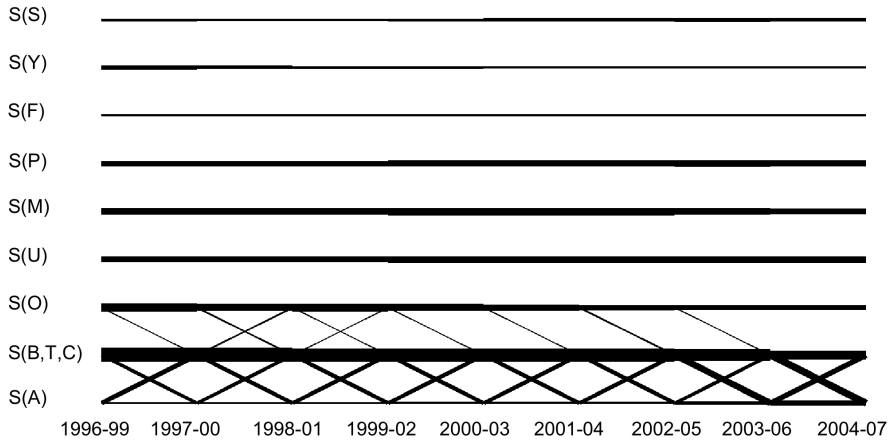


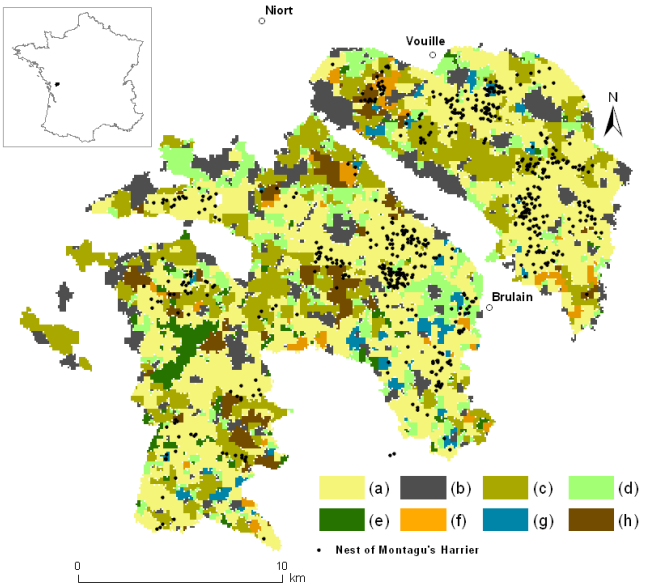
Figure 5. *Probabilités a posteriori des transitions entre classes de successions d'occupation du sol dans un modèle de type 2. L'axe des abscisses représente la période d'étude répartie en sous périodes de 4 ans. L'axe des ordonnées représente les classes de successions d'occupation du sol impliquant : poiS (S), raY-grass (Y), Forêts et friches (F), Prairies et luzernes (P), Maïs (M), Urbain (U), Orge d'hiver (O), Blé (B), Tournesol (T), Colza (C), Autres (A). L'épaisseur des lignes est proportionnelle à la fréquence des transitions. Les transitions horizontales indiquent une stabilité inter annuelle. Les transitions diagonales indiquent des changements inter annuels*

un HHMM2 ergodique (tous les états sont inter connectés comme le montre la figure 3). L'état possédant la plus forte probabilité a posteriori étiquette chaque point de la fractale. L'ensemble est rangé dans un shapefile qui représente la couche de classification telle qu'elle est visualisée dans la figure 6. Le croisement dans un SIG de cette information avec la couche d'information relative à la localisation des nids de busards permet d'identifier des zones à prospecter plus profondément. Grâce à cette carte, l'enquêteur a une meilleure connaissance du type des exploitations agricoles rencontrées, ce qui va l'aider à mieux comprendre l'influence des mesures agro-environnementales (MAE) adoptées dans les exploitations qui visent en partie au maintien de ces espèces patrimoniales.

4. Conclusion

Cette méthode de fouille de données modélise la mosaïque parcellaire d'un territoire agricole à l'aide d'un champ de Markov caché temporo spatial irrégulier. Nous avons adopté un point de vue markovien pour représenter les évolutions temporelles et spatiales des occupations agricoles dans un paysage soumis à des enjeux environnementaux en utilisant des HMM2 capables de modéliser des dépendances sur une plus longue échelle temporelle et spatiale. En étudiant l'entropie du sytème d'allocation

12 Colloque STIC et Environnement.



état	description des OCS
a	dominante de successions avec Tournesol, Blé et Colza (70%)
b	urbain et péri-urbain (60%)
c	dominante de successions avec Maïs (60%)
d	dominante de Prairies (50%)
e	dominante de Forêts (70%)
•	emplacement de nid de busard de Montagu

Figure 6. *Carte des zones homogènes en termes des successions d'occupation du sol d'un territoire dans la Plaine de Niort durant la période 1996-2007. Les emplacements de nids de busards de Montagu sont représentés et déterminent des zones d'intérêt dans lesquelles les exploitations agricoles seront enquêtées pour évaluer l'influence des mesures agro-environnementales (MAE)*

des terres, nous avons développé des méthodes pour fixer les paramètres temporels et spatiaux de ce champ. Le système de voisinage irrégulier de la mosaïque a été pris en compte par un parcours du plan par une courbe fractale dont la profondeur s'ajuste à la taille de la micro-parcelle. Cette information permet de ne représenter que par un point toutes les portions du territoire qui n'ont hébergé qu'une succession de cultures pendant la période d'étude. Cette modélisation a permis la localisation des régularités temporelles et spatiales dans les successions d'occupation du sol et de les croiser avec l'emplacement des nids de busards afin de mieux connaître les zones de prédilection de ces animaux. Nous envisageons maintenant de nous intéresser aux voisinages des parcelles afin d'identifier les quartiers de cultures plus aptes à accueillir le busard de

Montagu et à fouiller leur évolution. Ce travail de fouille de données temporelles et spatiales répondant à des préoccupations écologiques et agronomiques est un travail long et multi disciplinaire qui fait collaborer des informaticiens qui adaptent et testent des modèles numériques de représentation de territoires, des agronomes qui cherchent à comprendre les logiques des agriculteurs à partir des résultats de la fouille et des écologues soucieux d'aménager l'habitat des espèces à protéger au sein de territoires agricoles soumis à des contraintes économiques.

Remerciements

Nous remercions la région Lorraine, l'ANR BiodivAgrim, et l'API Ecoger pour leurs financements.

5. Bibliographie

[BAR 00] BARKER S., RAYNER P., « Unsupervised Image Segmentation Using Markov Random Field Models », *Pattern Recognition*, vol. 33, 2000, p. 587 – 602.

[BAU 72] BAUM L. E., « An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes », *Inequalities*, vol. 3, 1972, p. 1 – 8.

[BEN 03] BENOÎT M., LE BER F., MARI J.-F., MIGNOLET C., SCHOTT C., « CarrotAge, un logiciel pour la fouille de données agricoles », *Colloque STIC et Environnement SE'2003, Rouen, France*, INSA Rouen, Jun 2003.

[BEN 95] BENMILOUD B., PIECZYNSKI W., « Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images », *Traitement du signal*, vol. 12, n° 5, 95, p. 433 – 454.

[CAS 08] CASTELLAZZI M., WOOD G., BURGESS P., MORRIS J., CONRAD K., PERRY J., « A systematic representation of crop rotations », *Agricultural Systems*, vol. 97, 2008, p. 26 – 33.

[DER 04] DERRODE S., CARINCOTTE C., BOURENNANE S., « Unsupervised Image Segmentation Based on High-Order Markov Chains », *IEEE Trans. ICASSP*, 2004, p. 769 – 772.

[GEM 84] GEMAN S., GEMAN D., « Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images », *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, 1984.

[GLE 92] GLENN-LEWIN D., VAN DER MAAREL E., « *Plant Succession : Theory and Prediction* », chapitre Patterns and processes of vegetation dynamics, p. 11 – 44, Chapman & Hall, London, 1992.

[JEN 01] JENSEN F., *Bayesian Networks and Decision Graphs*, Springer, 2001.

[JUL 75] JULIAN B., « Statistical Analysis of Non-lattice Data », *The Statistician*, vol. 24, n° 3, 1975, p. 179 – 195.

[LEB 06] LE BER F., BENOIT M., SCHOTT C., MARI J.-F., MIGNOLET C., « Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software », *Ecological Modelling*,

14 Colloque STIC et Environnement.

vol. 191, n° 1, 2006, p. 170 – 185, http ://hal.archives-ouvertes.fr/hal-00017169/fr/.

[MAR 02] MARI J.-F., LE BER F., BENOÎT M., « Segmentation temporelle et spatiale de données agricoles », *Actes des journées Cassini 2002, Crozon*, septembre 2002, p. 251–272.

[MAR 07] MARI J.-F., LARGOUET C., « *Raisonnements sur l'espace et le temps : des modèles aux applications* », chapitre 9 et 12, p. 249 – 271 et 305 – 316, Lavoisier, http ://hal.archives-ouvertes.fr/hal-00126022/fr/ édition, mars 2007, ISBN : 2-7462-1455-5.

[NOY 07] NOYON V., CLARAMUNT C., DEVOGELE D., « A Relative Representation of Trajectories in Geographical Spaces », *Geoinformatica, Springer*, vol. 14, n° 11, 2007, p. 479 – 496.

[P.T 99] P. THINON J. D., « Partage de l'espace rural pour la gestion des problèmes environnementaux et paysagers dans le Vexin Français », *cah. Agr.*, n° 8, 1999, p. 373 – 387.

Segmentation temporo-spatiale du bassin versant du Yar fondée sur des occupations du sol télédétectées

El Ghali Lazrak (1) ; Marc Benoît (1) ; Jean-François Mari (2)

(1) INRA, UR 055, SAD ASTER
Domaine du Joly, F-88500 Mirecourt
lazrak, benoit@mirecourt.inra.fr

(2) LORIA, UMR CNRS 7503 et INRIA-Grand Est
B.P. 239, F-54506 Vandœuvre-lès-Nancy
jfmari@loria.fr

(Article en cours de rédaction)

Introduction

Les modèles de Markov cachés d'ordre 2 (HMM2) ont montré leur intérêt dans la segmentation temporelle et (ou) spatiale de territoires à dominante agricole (MIGNOLET et al., 2007 ; LAZRAK et al., 2009). Dans ces travaux, les auteurs ont utilisé des bases de données d'OCS construites par des relevés de terrain annuels conduits sur du long terme impliquant une énorme quantité de travail humain.

Par ailleurs, les techniques d'observation satellitaires de la Terre nous fournissent des informations de qualité croissante pour améliorer notre regard sur les usages des terres mis en œuvre par les agriculteurs à des échelles de territoires régionaux. Ces informations nouvelles nécessitent de nouvelles techniques de traitement permettant de répondre aux questionnements d'agronomes.

En effet, les images de télédétection est une autre source permettant de construire des bases de données spatio-temporelles d'OCS. La qualité des images de télédétection et des post-traitements permettant d'affecter les pixels de l'image à des OCS a été grandement améliorée au cours des dernières années. Toutefois, les bases de données d'OCS issues de traitements d'images de télédétection présentent, des différences par rapport à celles issues de relevés de terrain. En particulier, les images de télédétection offrent une richesse thématique plus faible et n'identifient pas les limites parcellaires. Dans cet article, nous faisons l'hypothèse que notre méthode de modélisation présentée dans (LAZRAK et al., 2009) peut traiter des bases de données spatio-temporelles d'OCS issues de télédétection.

Nous présentons dans cet article une procédure de fouille de données spatio-temporelles d'OCS télédétectées que nous avons mis en œuvre sur le bassin versant du Yar (Bretagne, France) où l'activité agricole est à l'origine de problèmes d'eutrophisation dans la baie de Saint-Michel-en-Grève. Nous avons réalisé une segmentation temporo-spatiale sur le corpus de données d'OCS du bassin versant du Yar. Cette segmentation a permis de partitionner le bassin versant du Yar en zones homogènes d'évolution des OCS. Ce partitionnement décrit de manière simple l'Organisation Territoriale de l'Activité Agricole (OTAA) dans ce bassin versant et révèle des dynamiques cachées pouvant contribuer à mieux comprendre l'implication de l'OTAA dans l'eutrophisation de la baie de Saint-Michel-en-Grève.

Matériel et méthodes

La zone d'étude

Le bassin versant du Yar est situé à l'extrême nord-ouest du département des Côtes-d'Armor, en limite du département du Finistère (figure 4.1). C'est un bassin versant littoral d'une superficie de 61,5 km^2. Il est à l'origine de plus de 60 % des flux polluants alimentant la baie de Saint-Michel-en-Grève (CORGNE, 2004). Les cours d'eau de ce bassin versant apportent de grandes quantités d'éléments nutritifs sur le littoral et y entraînent un phénomène d'eutrophisation.

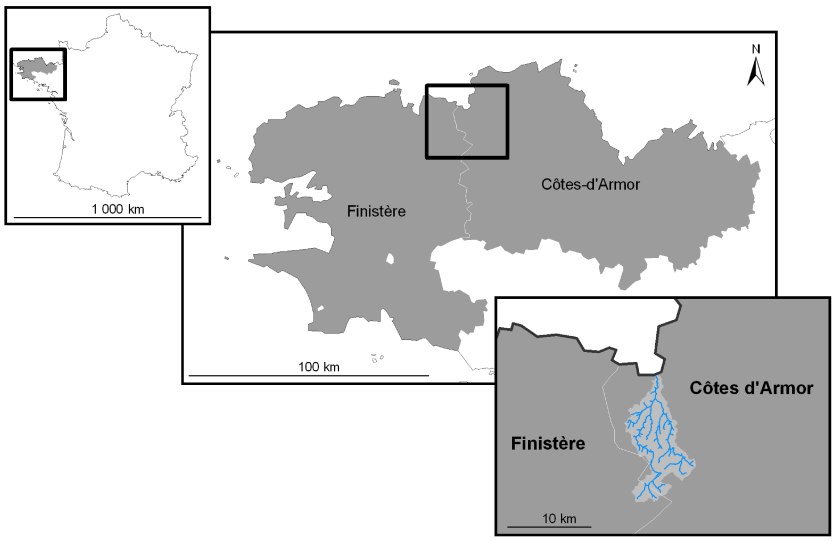


FIGURE 4.1 – Situation géographique du bassin versant du Yar. Le bassin versant du Yar est situé à l'extrême nord-ouest du département des Côtes-d'Armor, en limite du département du Finistère.

Détection des classes d'OCS

Les classes d'OCS ont été identifiées par télédétection puis traitement d'une série d'images satellites. Six OCS sont distinguées : Urbain (U), Eau (E), Forêt (F), Prairies (P), Céréales (C), Maïs (M). L'Urbain et la forêt sont issus de l'analyse de deux images satellites, une en début de la période d'étude (1997),

l'autre en 2003. Les autres OCS sont issues d'une identification annuelle à partir d'une série de 12 images prises en période estivale le long de la période d'étude : 1997-2008 (figure 4.2).

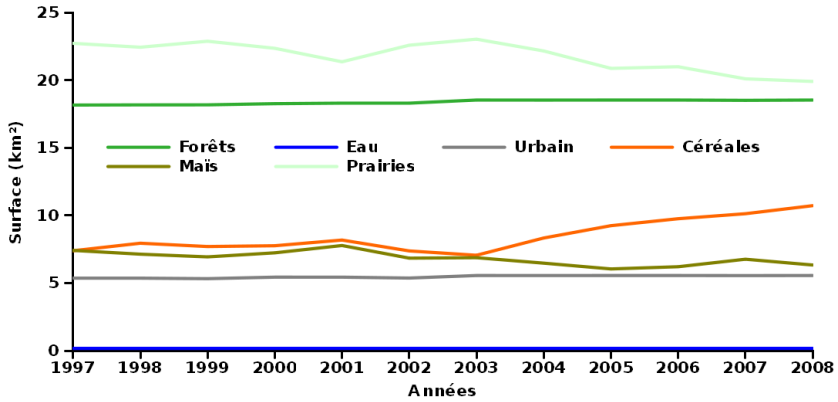


FIGURE 4.2 – Évolution des OCS dans le bassin versant du Yar durant la période d'étude

Le corpus d'OCS

Les données d'OCS issues du traitement des images satellites se présentent sous forme d'une couche vectorielle d'informations temporelles et spatiales stockée dans un système d'information géographique (SIG). Cette couche d'information vectorielle a été rastérisée moyennant un échantillonnage régulier avec une grille de points espacés de $20m \times 20m$. Le corpus résultant est une matrice où les colonnes représentent les OCS année par année et les lignes, les différents points d'échantillonnage localisés.

Définition d'un HMM2

Un HMM2 se définit par :

- un ensemble $E = \{e_1, e_2, \ldots, e_N\}$ à N états qui sont les résultats des variables X_t, où $t = 1, \ldots, T$,
- une matrice $A = (a_{ijk})$ des probabilités de transition entre les états définie sur E^3, où a_{ijk} est la probabilité *a priori* de transition $P(X_t = e_k / X_{t-2} = e_i, X_{t-1} = e_j)$ pour que le HMM2 soit dans l'état e_k dans l'indice t sachant qu'il était dans l'état e_j dans l'indice $t-1$ et e_i dans l'indice $t-2$ et
- un ensemble de N distributions discrètes : $b_i(\cdot)$ est la distribution des observations associées à l'état e_i. Cette distribution peut être paramétrique, non paramétrique ou bien être donnée par un HMM lorsqu'il s'agit d'un HMM hiérarchique (HHMM) (FINE et al., 1998 ; MARI et LE BER, 2006 ; LAZRAK et al., 2009).

ARPENTAGE : un logiciel de fouille de données basé sur les HMM2

L'ensemble des programmes permettant la spécification des HMM2, leur apprentissage à l'aide d'un corpus de données, la segmentation en zones homogènes représentées par les états cachés de l'HMM2

hiérarchique (que nous appelons super états ou états spatiaux) ainsi que la création des shapefiles pour visualiser ces zones constituent la boîte à outils ARPENTAGE (Analyse de Régularités dans les Paysages : Environnement, Territoires et Agronomie).

Segmentation temporo-spatiale avec un HHMM2

La fouille de données temporo-spatiale d'OCS avec des HHMM2 est un moyen pratique pour construire des modèles stochastiques capturant la variabilité temporelle et spatiale des données (LAZRAK et al., 2009). Nous modélisons la structure spatiale du bassin versant du Yar par un champ de Markov dont les sites sont des successions d'OCS. Ceci conduit à la définition d'un HHMM2 (figure 4.3) où un HMM2 maître ergodique approxime le champ de Markov. Chaque super état du HMM2 maître est associé à un HMM2 temporel modélisant la variabilité temporelle des successions d'OCS comme décrit par MARI et LE BER (2006) et LAZRAK et al. (2009). Ce HHMM2 est utilisé pour segmenter le bassin versant du Yar en zones homogènes en termes de l'évolution des distributions des OCS.

Pour la segmentation du bassin versant du Yar en zones homogènes d'évolution des OCS, des essais préliminaires de segmentation avec des HHMM2 au nombre croissant d'états spatiaux nous ont conduit à opter pour une segmentation en 5 super états (HHMM2 à 5 états spatiaux), les HMM2 temporels sont fixés à 12 états, un état par année.

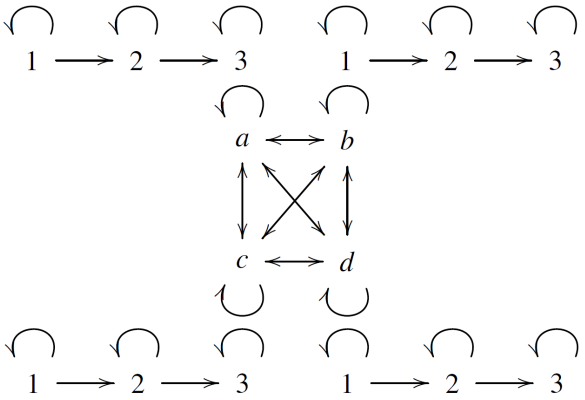


FIGURE 4.3 – HHMM2 utilisé pour fouiller le corpus spatio-temporel d'OCS du bassin versant du Yar. Chaque super état a, b, c, d du HMM2 maître (modèle ergodique) est un HMM2 temporel (modèle linéaire) dont les états sont 1, 2, 3. Tous les états ne sont pas représentés. Le HHMM2 utilisé pour modéliser l'OTAA dans le bassin versant du Yar compte 5 super états dont chacun est associé à un HMM2 linéaire de 12 états.

Résultats

Dans la figure 4.4, le bassin versant du Yar est représenté par une mosaïque de patches d'évolution des OCS. Ces patches appartiennent à l'un des 5 super états du HHMM2.

Les états 1 et 2, représentant respectivement la Forêt et l'Urbain, se caractérisent par une stabilité durant la période d'étude. La légère variation en 2003 contrastée par la stabilité les autres années est due

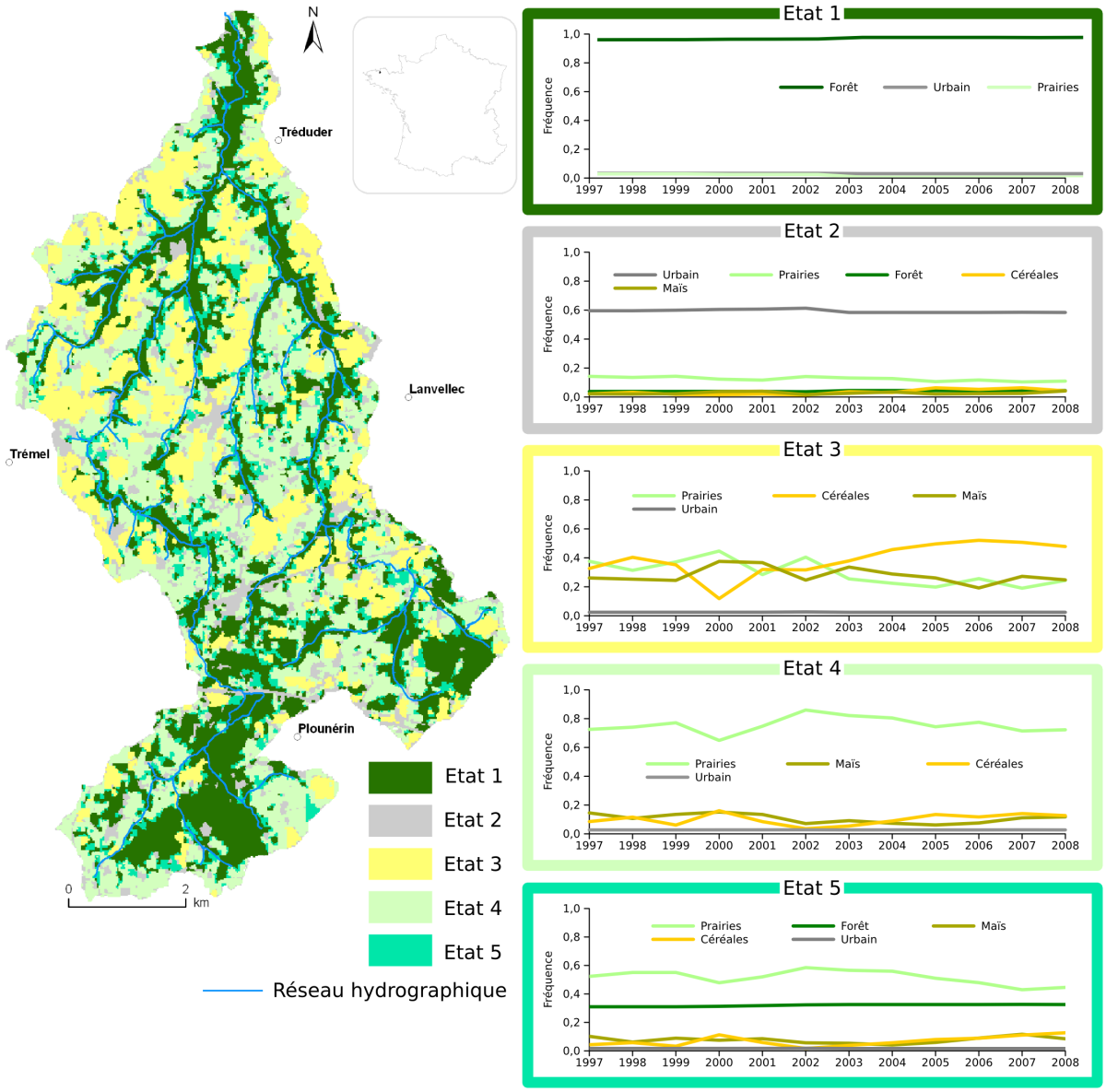


FIGURE 4.4 – Carte du bassin versant du Yar segmenté en zones homogènes d'évolution des OCS. Chaque unité cartographique correspond à un super état du HHMM2 qui a servi à la segmentation. Chaque super état est décrit par un diagramme d'évolution des distributions des fréquences des OCS. Le cadrant en haut à droite de la carte situe, par une tache noire, le bassin versant du Yar dans la carte de la France.

à la nature de la source d'information à l'origine de ces 2 OCS. En effet, ces OCS sont issues de l'analyse de deux images satellites, une datant du début de la période d'étude (1997) et l'autre de 2003.

Les 3 autres états exhibent une plus grande diversité d'OCS ainsi qu'une variation temporelle plus prononcée. Dans l'état 3 coexistent Pairies, Maïs et Céréales. Ces 3 OCS évoluent, jusqu'en 2002 en dents de scie autour d'une fourchette de fréquence entre 0,2 et 0,4. A partir de 2003, Prairies et Maïs régressent et sont remplacés par les Céréales dont la fréquence a nettement augmenté. Un changement du système de culture aurait été opéré dans les patches appartenant à l'état 3. Après 2003, cet état semble devenir caractéristique d'un système de culture où Prairies et Maïs deviennent interchangeables.

L'état 4 est principalement un état de Prairies où coexistent Maïs et Céréales avec de faibles fréquences.

L'état 5 est un état proche de l'état 4 où prédominent les Prairies et où Céréales et Maïs évoluent conjointement avec un même niveau de fréquence. Mais il s'en distingue par la proximité des patches le composant vis-à-vis des Forêts et par une tendance plus marquée de la régression des Prairies.

L'année 2000 est marquée par une forte variation de fréquences des Prairies et Céréales dans les états 3, 4 et 5. Cette variation est d'une même allure dans les états 4 et 5 alors que les évolutions sont inversées dans l'état 3. Ceci laisse supposer une différence dans les systèmes de cultures pratiqués dans l'état 3 d'une part et dans les états 4 et 5 d'autre part.

L'analyse des résultats de cette segmentation révèle une liaison stable par tranches de périodes entre Prairies, Céréales et Maïs. Cette liaison est ponctuellement rompue (2000) puis profondément modifiée (2003).

Discussion et conclusion

Bien que les OCS télédétectées soient thématiquement moins riches que celles issues des relevés de terrain (LAZRAK et al., 2009 ; MIGNOLET et al., 2007), la segmentation temporo-spatiale avec les HHMM2 a réussi à révéler des dynamiques dans les systèmes de cultures pratiqués dans ce bassin versant. Ces dynamiques pourraient contribuer à mieux comprendre l'implication de l'OTAA dans l'eutrophisation de la baie de Saint-Michel-en-Grève. Nous avons identifié des zones agricoles où les prairies temporaires seraient pratiquées en rotation avec le maïs et les céréales. Ces zones ont été le plus marquées par le changement des systèmes agricoles amorcé en 2003. L'interprétation des résultats de cette première fouille exploratoire sont à considérer comme des hypothèses qu'il conviendrait de vérifier avec des experts (agronomes de terrain, hydrologues, ...) ayant une bonne connaissance du bassin versant. D'autres besoins de fouille peuvent apparaître suite à une collaboration avec ces experts. Citons, par exemple, la recherche des successions affectées par la rupture survenue en 2003, ou l'analyse des dépendances des voisinages entre les successions d'OCS pour répondre à une question du genre : quel effet exerce la proximité des champs cultivés vis-à-vis des forêts ou de la zone urbaine ?

Références

CORGNE, S (2004). « Hiérarchisation des facteurs de changements de l'occupation hivernale des sols: Application au bassin versant du Yar (Bretagne) ». Dans : *Norois* 193, p. 17–29.

FINE, S, Y SINGER et N TISHBY (1998). « The Hierarchical Hidden Markov Model: Analysis and Applications ». Dans : *Machine Learning* 32, p. 41–62.

LAZRAK, EG, J-F MARI et M BENOÎT (2009). « Landscape regularity modelling for environmental challenges in agriculture ». Dans : *Landscape Ecology* 25.2, p. 169–183.

MARI, J-F et F LE BER (2006). « Temporal and Spatial Data Mining with Second-Order Hidden Markov Models ». Dans : *Soft Computing*. ISSN:1432-7643 10.5, p. 406–414.

MIGNOLET, C, C SCHOTT et M BENOÎT (2007). « Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale ». Dans : *Science of the Total Environment* 375.1-3, p. 13–32.

Chapitre 5

Analyse des relations de voisinage des successions d'OCS

Dans ce chapitre nous présentons notre procédure de fouille de données permettant d'analyser les dépendances entre les successions d'OCS voisines à l'échelle d'un territoire agricole de dimension régionale. Nous posons l'hypothèse que chaque agriculteur tient compte du voisinage lorsqu'il alloue les cultures à chacune de ses parcelles. En effet, nous faisons l'hypothèse que les agriculteurs organisent leurs assolements en quartiers culturaux où ils font côtoyer dans l'espace les cultures qui forment une succession dans le temps. À l'intérieur de ces quartiers culturaux, la variabilité spatio-temporelle des cultures n'est qu'apparente car elle concerne un même système de culture. Réduire cette variabilité pour ne garder que les variations significatives des systèmes de cultures permet de décrire l'OTAA de façon simple et efficace mais ceci nécessite au préalable de comprendre comment une succession d'OCS dans une parcelle dépend des successions des parcelles voisines. Pour prendre en compte l'irrégularité du système de voisinage de la mosaïque agricole, nous avons calculé un parcours de la zone d'étude à l'aide d'une courbe fractale de Peano à profondeur variable. Nous avons appliqué cette procédure de fouille au site d'étude de Chizé à travers deux articles :

Le premier article (LAZRAK et al., 2010) propose de modéliser la variabilité temporelle et spatiale de la mosaïque agricole. Cet article analyse les dépendances entre parcelles et permet l'identification de quartiers culturaux au sein desquels les occupations des parcelles manifestent de forts degrés d'attraction.

Le deuxième article (MARI et al., 2010) explore la complémentarité de deux approches d'analyse des relations de voisinages entre successions d'OCS dans un territoire agricole représenté par sa mosaïque de parcelles. La première approche, décrite dans (LAZRAK et al., 2009), segmente la mosaïque agricole en patches et représente ces patches à l'aide de deux distributions : une distribution qui représente l'homogénéité spatiale en termes des successions d'OCS et l'autre, l'influence des voisinages à la lisière des patches. La deuxième approche, décrite dans (LAZRAK et al., 2010), identifie l'organisation des systèmes de cultures en quartiers culturaux formés de parcelles dans lesquelles les mêmes rotations sont

pratiquées. Les deux approches se confortent l'une l'autre et permettent de modéliser les relations de voisinages entre successions d'OCS dans un territoire agricole de taille régionale.

Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales

Lazrak, El Ghali (1) ; Benoît, Marc (1) ; Mari, Jean-François (1, 2)

(1) INRA, UR 055, SAD ASTER

Domaine du Joly, F-88500 Mirecourt

{lazrak, benoit}@mirecourt.inra.fr

(2) LORIA, UMR CNRS 7503 et INRIA-Grand Est

B.P. 239, F-54506 Vandœuvre-lès-Nancy

jfmari@loria.fr

Abstract: The agricultural landscape can be seen as an assemblage of farm territories. The way farmers organize these territories is a time AND spatial process. Understanding how a land-use succession (LUS) in a parcel depends on LUS of the neighbouring parcels is a milestone to understand the time-spatial organization of the landscape mosaic. In this work, we analyse these time-space dependencies at agricultural landscape scales. We have performed a data mining process based on hidden Markov models (HMM) to identify spatial clusters of similar distributions of LUS in 2 neighbouring parcels, furthermore called cliques. We applied this data mining process to a land-use data set covering the period from 1996 to 2007 of a 350 km² agricultural landscape located within the Niort Plain (France). To take into account the irregular neighbour system of the parcel mosaic, we used a variable depth Hilbert-Peano scan of the area covering the landscape. Through illustrative examples of two contrasted spatial stochastic clusters, we show that considering temporal cliques gives valuable information on the neighbour system in terms of attraction between LUS.

Keywords: HMM2, data mining, temporal cliques

Introduction

In agricultural landscapes, land-uses are heterogeneously distributed among different agricultural parcels designed by farmers. At a first glance, the landscape spatial organization and its temporal evolution seem both random. Nevertheless, they reveal the presence of logical processes and driving forces related to the soil, climate, cropping system, and economical pressure. The mosaic of parcels together with their soil-occupancies (OCS) can be seen as a noisy picture generated by these different processes. The understanding of how the temporal succession of a parcel influences the neighbouring parcels is a milestone in the data mining process that aims at extracting knowledge from this mosaic. Furthermore, this piece of knowledge is helpful to simulate coherent agricultural landscapes (Le Ber et al., 2009). Recent studies (Le Ber et al., 2006 ; Castellazzi et al., 2008) have shown that the ordered sequences of OCS in each field can be adequately modelled by a Markov process. The OCS at time t depends upon the former OCS at previous times: $t-1$, $t-2$, ...
The Markov model or the hidden Markov model (HMM) are able to capture a limited amount of the temporal variability and allow the specification of land-use successions (LUS) in term of which the agricultural landscapes can be described in a more simple way (Lazrak et al., 2009). Similarly, in the spatial domain, the stochastic modelling of situated observations such as OCS or LUS by means of Markov fields is an elegant way to cluster a landscape into homogeneous patches described by probabilistic distributions of the situated observations.
In this work, we process at the same level the temporal and spatial information given by the parcels and their OCS and consider a pair of OCS in 2 neighbouring parcels at time slots t – furthermore called a temporal clique – rather than a single OCS as the basic temporal and spatial information. The stochastic modelling of the temporal cliques allows a spatial and temporal clustering of the landscape and gives valuable informations on the time and spatial dependencies between OCS. Our objective is to develop a generic data mining process, based on HMM and temporal cliques, in order to highlight these time-space dependencies at agricultural landscape scales.

The land-use database

The case study area is a 350 km² agricultural landscape located within the Niort Plain in Poitou-Charentes region, France. This agricultural landscape has been surveyed for more than 12 years (1996 – 2007). Every year, two land-use surveys (in April and June) allow to monitor both early harvested and late planted crops. These surveys are stored in a GIS geodatabase, in a vector format.
An analysis based on the average frequency of land-uses over the 12-year study period reveals 47 land-uses. These land-uses have been grouped with the help of agricultural experts in 10 categories (table 1) following an approach based on the similarity of crop management.

Table 1. Composition and average frequencies of adopted land-use categories (Lazrak et al., 2009)

Land-use category	Land-use	Cumul. frequency
Wheat	Wheat, bearded wheat, cereal	0.337
Sunflower	Sunflower, reygrass followed by sunflower	0.476
Rapeseed	Rapeseed	0.600
Urban	Built area, peri-village, road	0.696
Grassland	Grassland of various types, alfalfa, ...	0.774
Maize	Maize, ryegrass followed by maize	0.850
Forest	Forest or hedge, wasteland	0.884
Winter barley	Winter barley	0.918
Ryegrass	Ryegrass, ryegrass followed by ryegrass	0.942
Pea	Pea	0.964
Others	Spring barley, grape vine, clover, field bean, ryegrass, cereal-legume mixture, garden/market gardening, ...	1.000

The agricultural landscape mosaic

The agricultural landscape can be seen as an assemblage of polygons of variable size – the parcels – where each parcel holds a given OCS.
A parcel can be bounded by a road, a path or a limit of a neighbouring parcel. The parcel boundaries can change every year. To take account of this change, each year, the surveyors update the edges – the boundaries – of parcels in the GIS geodatabase. This led to the definition of the elementary parcel as the result of the spatial union of previous parcel edges (figure 1). There are about 20,000 elementary parcels in the study area over the 1996 – 2007 period. Each elementary parcel holds one succession of OCS during the study period.
The corpus of land-use data is sampled using a regular grid and is represented in a matrix in which the rows represent the land-uses year by year and the lines, the different grid locations.

Cliques and temporal cliques

Two elementary agricultural parcels represented by 2 polygons are neighbouring if they have at least an edge in common. A clique is a set of parcels in which two unspecified parcels are neighbour. In the mosaic of polygons, the neighbouring relationship – called the neighbour system – is irregular. The parcels have a variable number of neighbours in different geographical directions as opposite to digital images where a site has a fixed number of neighbours. In this paper, we consider simple cliques made of 2 neighbouring parcels represented by the 2 centroids of the parcels. Experimental preliminary results show that the OCS distribution in the cliques is isotropic: the direction defined by the 2 centroids does not carry any information.

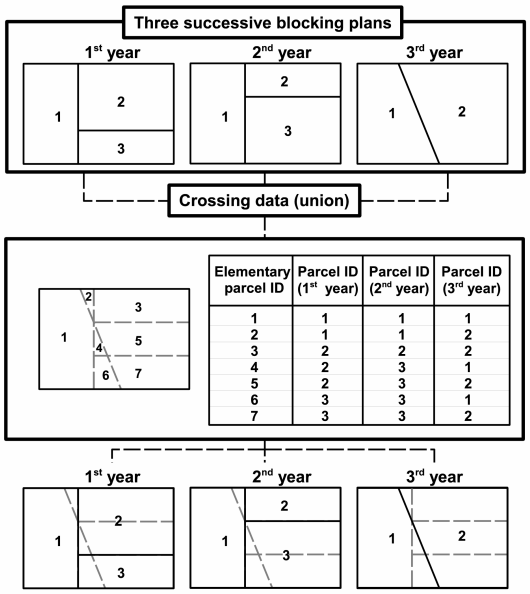


Figure 1. An example of parcel boundary evolution over three successive years. The union of parcel boundaries during this period leads to the definition of seven elementary parcels

Following Benmiloud and Pieczynski (Benmiloud and Pieczynski, 1995 ; Pieczynski, 2003), we have approximated the Markov field by scanning the 2-D landscape representation with a Hilbert-Peano curve (figure 2). The Markov field is then represented by a Markov chain. To take into account the irregular neighbour system, we have first regularly sampled the area covering the landscape (eg.1 point every 20 m), next have introduced an Hilbert-Peano scanning and finally, have adjusted the fractal depth to the elementary parcel size. The figure 2 illustrates this concept. The sites lying in the same elementary parcel are agglomerated into one point as far they draw the fractal motif. Two successive sites in the L-Length fractal curve (s_{l-1}, s_l), $1 \leq l < L$, define a clique.

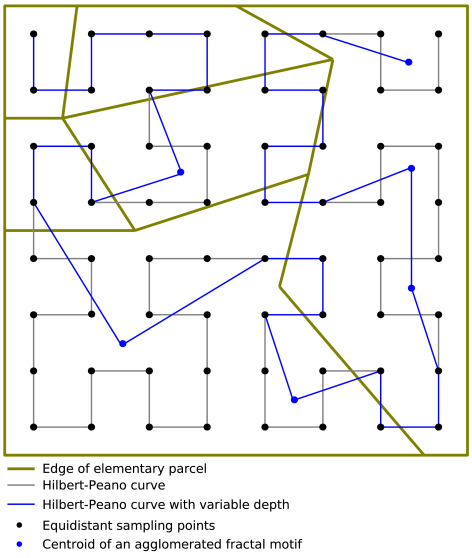


Figure 2. Variable depth Hilbert-Peano scan to take into account the parcel size. Two successive merging in the bottom left parcel yield to the agglomeration of 16 points

This scanning introduces a spatial warping and a surface normalization in the parcel mosaic. Large parcels are less sampled, whereas no site agglomeration occurs when the curve crosses the parcel boundaries. The longer is the boundary between two polygons, the more frequent is the clique. Of course, the parcels having singular shape cannot be represented with one centroid and some cliques are situated into the same elementary parcel (figure 2). As a matter of fact, the problem of visiting only once the edges or the vertexes of a graph is known to be NP (non polynomial) hard: there is no algorithm running in a reasonable time to solve it (Rubin, 1974). Our irregular spatial sampling is a crude way to avoid this issue.

The occupancies of a site and its neighbour at time t define the temporal clique. At each site s_l in the variable depth Hilbert-Peano scan, we have defined a feature vector o_l^t with the OCS held in the (s_{l-1}, s_l) cliques:

$$o_l^t = ((s_{l-1}^t, s_l^t), (s_{l-1}^{t+1}, s_l^{t+1})) \ , \qquad 0 \le t < T-1, \quad 1 \le l < L \tag{1}$$

where s_l^t is the OCS at time t and index l in the variable depth fractal curve. t is a time index running over the study period, and l the spatial index in the L-Length scanning curve. At time t, a landscape is then represented by a $(L-1)$-Length sequence of overlapping temporal cliques. We consider also $T-1$ representations to cover the T year length study period due to the overlap artefact.

The cliques inside the same elementary parcel result from the variable depth Hilbert-Peano scan. They are not interesting in the present study. To partially deal with this artefact, feature vectors o_l^t verifying $(s_{l-1}^t = s_l^t)$ and $(s_{l-1}^{t+1} = s_l^{t+1})$ are removed from the resulted distributions.

The feature vector o_l^t is the outcome of 4 random variables S_t, N_t, S_{t+1} and N_{t+1} that define the observable stochastic process (*cf.* table 2 and table 3).

The time-space Markovian modelling framework

The way a farmer organizes his territory is a time and spatial process. This time-space dependency becomes more complex at agricultural landscape scales when the agricultural mosaic is built under many farmer's logics. To analyze these dependencies, we rely on 2 assumptions:

1. the OCS of a given field depends upon the OCS of the neighbouring fields (the MRF assumption), and
2. the OCS of a given field in a given year depends also upon the OCS of recent previous years (the Markov chain assumption).

We have modelled the spatial structure of the landscape by a MRF whose sites are random variables of temporal cliques. Like in our previous works (Mari and Le Ber, 2006 ; Lazrak et al., 2009), the MRF has been approximated by a HMM2. This HMM2 has been trained by the EM algorithm on the $T-1$ temporal representations of the landscape.

The time-space clustering

The stochastic modelling and clustering exhibits patches characterized by distributions of temporal cliques.

• The analysis of rows S_t and S_{t+1} shows the time dependencies at the site level whereas the analysis of rows N_t and N_{t+1} shows the same time dependencies at the neighbour level;

• similarly, the analysis of rows S_t and N_t shows the attraction between OCS;

• furthermore, the joint analysis permits to quantify the attraction between LUS.

Table 2 is a simple example involving the patches tagged as Urban by the stochastic clustering. We can see that the Grassland and Urban categories are stable in the time and have a mutual strong attraction. Less frequent is the neighbourhood occupied by crop successions involving Wheat, Rapeseed and Sunflower.

Table 2. Temporal cliques in the patches tagged as Urban by the stochastic clustering. Items are listed in decreasing order of frequency.

S_t	N_t	S_{t+1}	N_{t+1}
Urban	Grassland	Urban	Grassland
Grassland	Urban	Grassland	Urban
Sunflower	Urban	Wheat	Urban
Urban	Sunflower	Urban	Wheat
Urban	Wheat	Urban	Rapeseed

Table 3. Temporal cliques in the spatial cluster holding crop successions including Sunflower, Wheat, and Rapeseed. Items are listed in decreasing order of frequency.

S_t	N_t	S_{t+1}	N_{t+1}
Wheat	Rapeseed	Rapeseed	Wheat
Rapeseed	Wheat	Wheat	Rapeseed
Sunflower	Rapeseed	Wheat	Wheat
Rapeseed	Sunflower	Wheat	Wheat
Wheat	Wheat	Rapeseed	Sunflower
Wheat	Wheat	Sunflower	Rapeseed
Sunflower	Wheat	Wheat	Rapeseed
Wheat	Sunflower	Rapeseed	Wheat
Rapeseed	Wheat	Wheat	Sunflower
Wheat	Rapeseed	Sunflower	Wheat

The table 3 is an other example that represents the most frequent items of temporal cliques in the patches holding crop successions including Sunflower, Wheat, and Rapeseed. This table shows clearly that, in theses patches, the OCS located nearby a parcel will be held soon in this parcel. Most likely, this time-space relationship is dictated by the type of crop rotations practiced in this cluster. In fact, a previous data mining study (Lazrak et al., 2009) on the same land-use data base allowed to discover that the main rotations involving Sunflower, Wheat, and Rapeseed in the study area are the quadrennial rotation: (Sunflower-Wheat-Rapeseed-Wheat), and the biennial rotations: (Sunflower-Wheat) and (Rapeseed-Wheat). Furthermore, this spatial cluster describes an open-field agricultural area because the temporal cliques involving either Forest or Grassland in the neighbourhood are not represented.

Discussion

We have proposed a new representation of agricultural landscapes based on temporal cliques of parcels. To cope with the irregular neighbour system between the parcels,

we have specified a variable depth fractal curve that introduces a surface normalization factor and visits the parcels according to their neighbourhood. The sampling becomes irregular and enhances the neighbourhood effects.
Considering temporal cliques rather than single OCS gives a valuable information about the neighbour system between OCS and LUS. This shows the different degree of attraction between LUS in this area and therefore describes the landscape through patches.
Compared to our previous work (Lazrak et al., 2009), the stochastic modelling of the parcel mosaic based on temporal cliques clusters a landscape into agricultural districts that reveal the LUS and the LUS attraction. We put forward the hypothesis that these agricultural districts capture the temporal and spatial variability and can describe, in a simpler way, the agricultural landscapes to achieve a better understanding of the underlying logical processes.

Acknowledgments

This work was supported by the ANR-ADD-COPT project, the API-ECOGER project and the ANR-BiodivAgrim project. We thank the CNRS team in Chizé for their data records obtained from their "Niort Plain data base".

References

Benmiloud B., Pieczynski W., 1995. Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images, *Traitement du signal*, 12(5), p. 433 – 454.
Castellazzi M., Wood G., Burgess P., Morris J., Conrad K.F., Perry J.N., 2008. A systematic representation of crop rotations, *Agricultural Systems*, 97, p. 26–33.
Lazrak E.G., Mari J.-F., Benoît M., 2009. Landscape regularity modelling for environmental challenges in agriculture, *Landscape Ecology*, Sept. 2009. http://hal.inria.fr/inria-00419952/en/.
Le Ber F., Benoit M., Schott C., Mari J.-F., Mignolet C., 2006. Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software, *Ecological Modelling*, 191(1), Jan 2006, p. 170 – 185. http://hal.archives-ouvertes.fr/hal-00017169/fr/.
Le Ber F., Lavigne C., Adamczyk K., Angevin F., Colbach N., Mari J.-F., Monod H., 2009. Neutral modelling of agricultural landscapes by tessellation methods – application for gene flow simulation, *Ecological Modelling*. http://hal.archives-ouvertes.fr/hal-00409081/fr/.
Mari J.-F., Le Ber F., 2006. Temporal and Spatial Data Mining with Second-Order Hidden Markov Models, *Soft Computing*, 10(5), March 2006, p. 406 – 414. http://hal.inria.fr/inria-00000197.
Pieczynski W., 2003. Markov models in image processing, *Traitement du signal*, 20(3):255–278.
Rubin F., 1974. A search procedure for hamilton paths and circuits, *Journal of the ACM*, 21, Oct. ISSN 0004-5411.

Fouille de paysages agricoles : analyse des voisinages des successions d'occupation du sol

Jean-François Mari[1] El Ghali Lazrak[2] Marc Benoît[2]

[1] LORIA, UMR CNRS 7503 et INRIA-Grand Est
[2] INRA, UR 055, SAD ASTER

LORIA, UMR CNRS 7503 et INRIA-Grand Est
B.P. 239, F-54506 Vandœuvre-lès-Nancy
jfmari@loria.fr
INRA, UR 055, SAD ASTER
domaine du Joly, F-88500 Mirecourt
{lazrak, benoit}@mirecourt.inra.fr

Résumé

Nous présentons deux approches stochastiques à l'aide de modèles de Markov cachés (HMM) pour décrire les relations de voisinages entre successions d'occupation du sol dans un paysage agricole représenté par sa mosaïque de parcelles. Une première approche qualifiée de temporo-spatiale recherche à l'aide d'un HMM des classes de successions d'occupation du sol qui sont par la suite localisées. Ces dernières s'agglomèrent au sein de régions compactes (ou patches*). Nous présentons une méthode de clustering spatial qui représente les patches à l'aide de deux distributions : une distribution qui représente l'homogénéité spatiale et l'autre, l'influence des voisinages à la lisière des patches. La deuxième approche analyse les dépendances spatio-temporelles représentées par le graphe du système de voisinages entre parcelles et enrichit les connaissances sur l'organisation territoriale de l'activité agricole en permettant l'identification de quartiers culturaux au sein desquels les occupations des parcelles manifestent de forts degrés d'attraction.*

Mots Clef

HMM2, classification spatiale, classification temporelle, successions de culture.

Abstract

We present two stochastic approaches using Hidden Markov Models to describe the relations between the neighborhoods of land use successions in a territory represented as a parcel mosaic. The first approach, qualified as time-spatial, clusters first the land use successions and, next, locates them in the territory. They appear agglomerated into compact areas called patches. We describe a clustering method that describes the patches by means of 2 distributions of land use successions : one to model the spatial homogeneity and the other to model the influence of the neighborhood. The second approach analyzes the time spatial dependencies in the neighborhood graph and gives valuable informations in defining cultural districts in which the land use successions show a high degree of attractivity.

Keywords

HMM2, temporal clustering, spatial clustering, land use successions

1 Introduction

Le paysage agricole peut être perçu comme un assemblage de polygones de tailles différentes – les parcelles – où chaque parcelle porte une occupation du sol (OCS). L'ensemble de ces OCS est choisi par les agriculteurs.
La façon dont chaque agriculteur organise son territoire est un processus à la fois temporel et spatial. Les dépendances entre l'espace et le temps deviennent complexes à l'échelle des paysages agricoles où la mosaïque agricole est construite par de nombreux acteurs. Ceux-ci sont soumis à des opportunités ou des contraintes de nature différente agissant à diverses échelles. En allant du plus proche au plus lointain, on peut distinguer les contraintes / opportunités de voisinage et de localisation par rapport à l'eau ou aux villages, celles d'ordre pédologique et topographique, celles imposées par la présence d'une coopérative et enfin celles définies par la politique agricole à une échelle supra nationale. L'ensemble de ces contraintes / opportunités nous fait émettre l'hypothèse que l'activité des différents agriculteurs, sans être concertée, conduit à l'aménagement du paysage d'une manière convergente et intelligente qu'il nous faut comprendre avant de pouvoir la modifier.
Pour analyser les dépendances temporelles et spatiales

entre OCS, nous nous appuyons sur 2 hypothèses :

1. l'OCS d'une parcelle dépend de l'OCS des parcelles voisines (hypothèse de champs de Markov) ;
2. l'OCS d'une parcelle une année donnée dépend de l'OCS de cette parcelle les années précédentes (hypothèse de chaîne de Markov).

Dans cet article, nous présentons une analyse des voisinages des successions par une approche stochastique avec des HMM d'ordre 2 (HMM2). Nous commençons par présenter le territoire d'étude et le corpus spatio-temporel d'OCS. Ensuite, nous présentons deux approches complémentaires d'analyse des voisinages des successions d'OCS : (i) une approche qualifiée de temporo-spatiale où il s'agit de rechercher des classes de successions d'OCS qui sont par la suite localisées et fait émerger des patches ; (ii) une deuxième approche qui envisage le couple (OCS, OCS d'une parcelle voisine) pour analyser les dépendances spatiales et temporelles entre OCS voisines en vue de valider et compléter les résultats de la première approche.

2 Représentation de la mosaïque parcellaire

2.1 Définitions

Le composant de base dans un territoire agricole est la parcelle : polygone de taille variable possédant un couvert : l'OCS. Elle est délimitée géographiquement par une route, un chemin, ... ou artificiellement par l'agriculteur qui subdivise le territoire de son exploitation pour respecter un assolement : ensemble des surfaces allouées à chaque culture. Les frontières des parcelles changent chaque année en fonction des choix des agriculteurs. Pour tenir compte de ce changement, les enquêteurs définissent l'ensemble des micro-parcelles comme étant l'union de toutes les frontières des parcelles pendant la période d'étude.
Le paysage agricole étudié s'étend sur 350 km^2 dans la Plaine de Niort. Il est enquêté depuis plus de 12 ans pour la localisation et les occupations de ses parcelles. Il y a environ 20000 micro-parcelles. Une micro-parcelle n'a hébergé qu'une succession de culture pendant la période d'étude. Les relevés d'OCS issus des enquêtes annuelles sont stockés dans un système d'information géographique sous format vectoriel et constituent une couche d'informations temporelles et spatiales. Cette couche d'information vectorielle a été rastérisée avec une grille de points régulièrement espacés (10m x 10m). Le corpus résultant est une matrice où les colonnes représentent les OCS année par année et les lignes, les différents points d'échantillonnage localisés. Le corpus compte au total 47 OCS que nous avons regroupées, dans un travail antérieur [2] (Tab. 1) suivant une démarche tenant compte de la similitude des conduites culturales. Nous considérons plus particulièrement 5 OCS appartenant à deux classes contrastées de couverts :
– les couverts dynamiques : Blé, Tournesol, Colza ;
– et les couverts pérennes : l'Urbain et les Prairies.

OCS	Fréq. cumulée
Blé (B)	0.337
Tournesol (T)	0.476
Colza (C)	0.600
Urbain (U)	0.696
Prairies et Luzernes (P)	0.774
Maïs (M)	0.850
Forêts et friches (F)	0.884
Orge d'hiver (O)	0.918
Ray-grass (Y)	0.942
Pois (S)	0.964
Autres (A)	1.000

TABLE 1 – Composition et fréquences moyennes sur le territoire des OCS adoptées

2.2 Choix de l'observation élémentaire

Nous envisageons plusieurs observations élémentaires qui constitueront les modalités d'un pixel de l'image représentant la mosaïque agricole :

1. l'OCS en un point d'une parcelle représentant son occupation ;
2. la succession d'OCS de longueur fixe (par exemple T-B-C-B) représentant la suite temporelle (4 ans dans notre exemple) des OCS en un point d'une parcelle ;
3. le couple (OCS, OCS d'une parcelle voisine).

La première observation est utile pour retrouver les rotations dominantes selon la méthode de fouille développée par [3, 5].
L'utilisation de successions d'OCS de longueur fixe permet de différencier les rotations de cultures sur des périodes différentes : par exemple les rotations bi, tri ou quadri-annuelles. Plus la succession est longue, plus elle intéresse l'agronome, mais plus le nombre d'observations sera important. Ce dernier conditionne la taille des distributions des états du HMM.
Enfin, l'utilisation de couples (OCS, OCS d'une parcelle voisine) permet de fouiller les voisinages entre cultures et leurs évolutions. Ceci permet de quantifier l'attraction spatiale des OCS comme le révèle la méthode décrite au §5.

2.3 Choix de la longueur de succession et de la résolution spatiale

L'exécution des algorithmes de classification dépend de deux facteurs : (i) la taille de la matrice des données à travers la résolution d'échantillonnage spatial qui joue sur le temps de calcul, (ii) et la longueur de la succession d'occupation du sol qui joue sur la place occupée par la représentation des distributions de probabilité. Afin de disposer d'un critère objectif pour le choix de la résolution spatiale, la perte d'information en termes de diversité des successions d'OCS a été quantifiée pour différentes résolutions d'échantillonnage spatial (figure 1-a). La résolution de 80 m x 80 m a été retenue. Cette résolution a permis d'obtenir une matrice de données 64 fois plus réduite que la ma-

trice de données initiale avec une perte de seulement 6% en termes de nombre de successions.
Nous nous sommes intéressés ensuite à l'influence de la longueur des successions sur l'entropie du système. Nous avons comparé le nombre de successions différentes à celles produites aléatoirement dans le même territoire. Chaque OCS est tirée selon une loi uniforme sans tenir compte ni de sa localisation ni des OCS précédentes au même endroit. La figure 1-b montre qu'à partir d'une longueur de succession de 4 années, la zone d'étude commence à se distinguer nettement du modèle où les successions sont choisies de manière aléatoire dans la zone d'étude. Ceci justifie notre choix des successions de 4 ans (appelées par la suite quadruplets) comme une observation élémentaire pertinente.

2.4 Échantillonnage spatial à résolution variable

Dans la mosaïque parcellaire, le système de voisinage est irrégulier. Une parcelle a un nombre quelconque de parcelles avec lesquelles elle partage une frontière commune. Nous modélisons la mosaïque parcellaire par un champ de Markov irrégulier dans lequel la parcelle s'inscrit dans une succession temporelle d'OCS. L'ensemble constitue une image de successions plutôt qu'une succession d'images et ne peut être étudié directement par les modèles numériques d'images appliqués au suivi de trajectoires d'objets [6]. Enfin, les territoires étudiés ne sont pas carrés et les parcelles ajoutées pour leur donner une forme carrée doivent être en nombre minimum afin de ne pas perturber la classification spatiale.
Pour transformer cette représentation 2D en représentation 1D afin d'utiliser des HMM plutôt que des champs de Markov, nous cherchons un parcours du graphe des voisinages. Comme ce problème est NP-difficile, nous avons adopté une solution approchée en définissant un parcours fractal sur la grille régulière et en adaptant la profondeur de fractalisation à la taille de la micro-parcelle. Un processus récursif fusionne les points situés sur la courbe quand ils correspondent au motif générateur et qu'ils sont dans la même micro-parcelle et donne naissance à une courbe fractale à profondeur variable appelée dans la suite *courbe fractale réduite* (cf. Fig. 2). Ce balayage particulier introduit une normalisation des surfaces des parcelles ; les grandes parcelles sont sous représentées alors que les points au voisinage des frontières entre parcelles sont sur représentés.

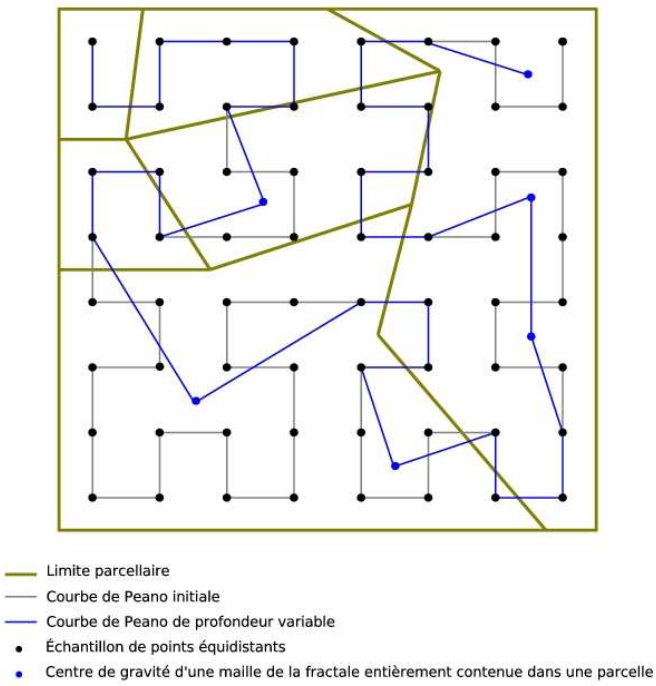


FIGURE 2 – Prise en compte de la taille des parcelles dans le parcours du plan par la courbe fractale de Hilbert-Peano. L'exemple montre deux fusions successives dans la parcelle située en bas à gauche aboutissant à l'agglomération de 16 points en un seul

3 Classification temporo-spatiale des successions

3.1 Classification temporelle des successions

Nous envisageons un HMM2 dont la topologie est donnée Fig.3 selon la méthode de fouille de données temporelles présentée dans [2]. Les successions d'OCS de chaque micro-parcelle sont utilisées dans l'algorithme Forward-Backward pour l'estimation du HMM2 [4] sans tenir compte de leur localisation. Les distributions initiales et finales des états sont données Tab. 2.
Les états du HMM2 ont été initialisés à l'aide de distributions uniformes de quadruplets d'OCS correspondant aux motifs de recherche décrits dans le tableau 2-a. Le HMM2 utilisé possède autant d'états que de S(X) et il est appris sur le corpus.
Dans le territoire d'étude, le Blé (B), le Tournesol (T) et le Colza (C) sont le plus souvent intégrés dans une même succession de 4 ans (p. ex. T-B-C-B, T-B-T-B, C-B-C-B, T-B-B-B, C-B-B-B). La classification avec une classe commune à ces trois cultures a permis d'obtenir des résultats plus cohérents qu'avec les classes séparées. Cette classe, notée S(B,T,C), regroupe les quadruplets impliquant au moins l'une de ces trois cultures. Avec ce regroupement, nous distinguerons par la suite 3 classes de successions : S(B,T,C), S(U) et S(P). Le tableau 2 résume la distribution des observations à l'intérieur de ces classes de successions après apprentissage du corpus d'étude sans prise en compte de la composante spatiale.

4 Spatialisation des S(X)

Les localisations des classes de successions d'OCS ont été recherchées à l'aide de modèles de Markov cachés hiérarchiques d'ordre 2 (HHMM2) [4]. Le HHMM2 utilisé est un modèle ergodique qui comporte un état spatial par classe

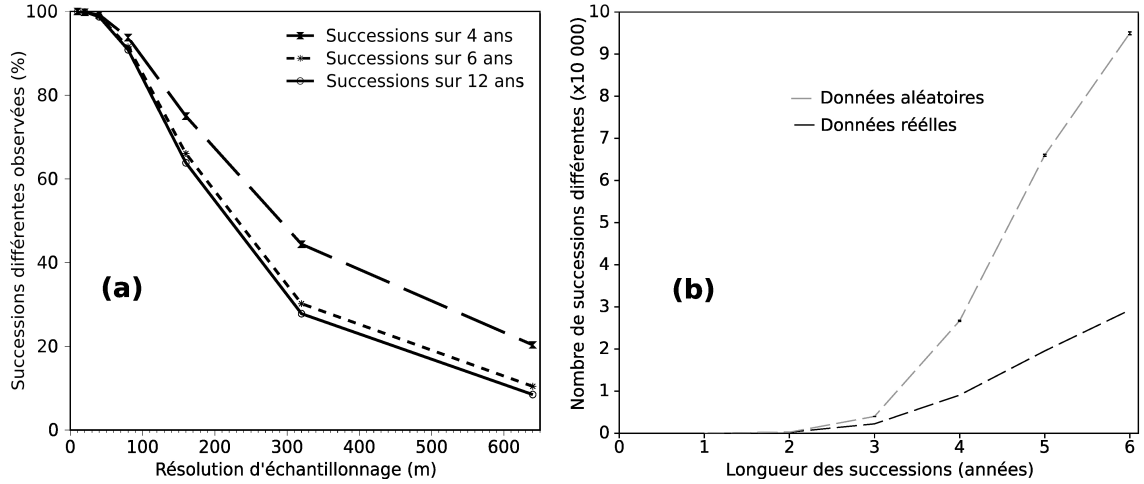


FIGURE 1 – (a) Évolution de la perte d'information spatiale en fonction de la résolution d'échantillonnage. (b) Comparaison de la diversité des successions d'OCS entre les données réelles et des données générées aléatoirement pour différentes longueurs de successions

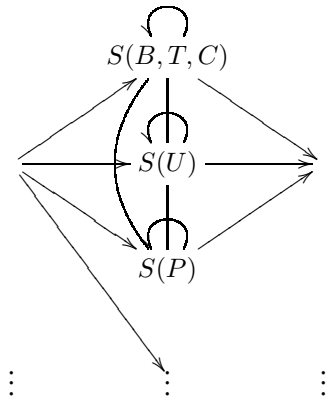


FIGURE 3 – HMM2 pour la classification temporelle des successions. Les successions d'OCS de chaque micro-parcelle estiment les états S(X) sans utiliser l'information de localisation. Les successions d'OCS sont représentées par des quadruplets d'OCS se chevauchant. Tous les états du HMM2 ne sont pas représentés

(a)

S(X)			
t	t+1	t+2	t+3
X	?	?	?
?	X	?	?
?	?	X	?
?	?	?	X

(b)

S(B,T,C)		S(U)		S(P)	
T-B-C-B	0.128	U-U-U-U	0.977	P-P-P-P	0.377
C-B-T-B	0.112	B-U-U-U	0.002	P-P-P-B	0.040
B-T-B-C	0.112	U-U-U-B	0.002	B-P-P-P	0.026
B-C-B-T	0.105	P-U-U-U	0.002	Y-P-P-P	0.026
T-B-T-B	0.072	T-B-U-U	0.001	T-P-P-P	0.025
C-B-C-B	0.065	U-U-B-U	0.001	B-T-P-P	0.018
B-C-B-C	0.062	U-U-B-U	0.001	P-P-P-T	0.016
B-T-B-T	0.059	T-U-U-U	0.001	P-P-T-B	0.015
B-C-B-B	0.033	P-P-U-U	0.001	A-P-P-P	0.015
B-B-C-B	0.024	U-U-U-P	0.001	P-P-B-C	0.013

TABLE 2 – Description des états du HMM2 utilisé pour la classification temporelle avant et après l'apprentissage. En (a) sont représentés les gabarits des quadruplets équiprobables impliquant l'OCS X. ? désigne une quelconque OCS. En (b) sont données les distributions des quadruplets dans les classes à l'issue de l'apprentissage sans prise en compte de l'espace

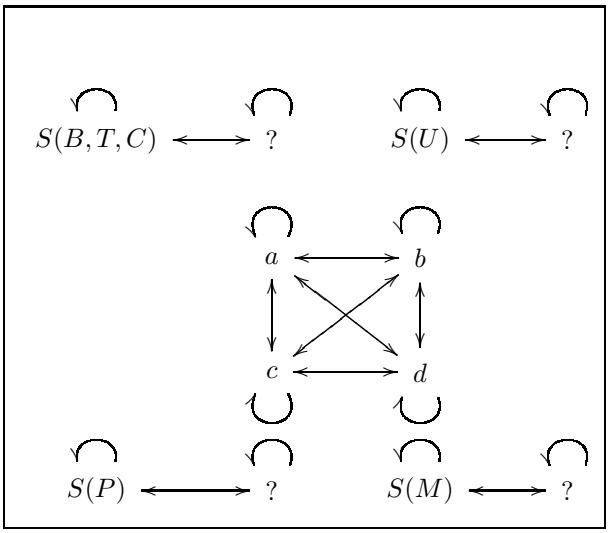


FIGURE 4 – Chaque état a, b, c, d du HHMM2 maître est un HMM2 temporel dont les états sont s(x) et un état équiprobable noté ? [4]. Tous les états du HHMM2 ne sont pas représentés

de successions à localiser comme le montre la figure 4. Chaque état spatial est défini par un HMM2 ergodique à 2 états : (i) un état de gabarit représentant la classe S(X) et initialisé par un apprentissage temporel tel qu'il est décrit dans le §3.1 et (ii) un état de réserve destiné à capturer les successions au voisinage ne répondant pas aux critères spécifiés dans l'état de gabarit. Cet état de réserve est appelé par la suite, état de voisinage. Le tableau 3 présente les classes de successions S(X) spatialisées avec leurs états de voisinages respectifs. Le fait d'analyser les successions dans l'ordre imposé par la courbe de Peano réduite (cf. Fig. 2) introduit un changement dans la distribution des quadruplets. Dans chaque état du HMM hiérarchique, l'état équiprobable capte les exceptions dues aux effets de voisinage. Un site n'est plus classé seulement en fonction de ses caractéristiques temporelles mais subit l'influence de ses voisins. La classe S(B,T,C) spatialisée montre une distribution de fréquences légèrement différente par rapport à la distribution de la même classe non spatialisée (Tab. 2). L'ordre des quadruplets a néanmoins été conservé dans les deux cas. L'état de voisinage accompagnant la classe S(B,T,C) spatialisée indique que les patches caractérisés par des successions impliquant majoritairement le Blé, le Tournesol et le Colza, comportent aussi des Prairies et luzernes (P-P-P-P) à 2% et des zones Urbaines (U-U-U-U) à 1%. La spatialisation des deux classes S(U) et S(P) a également modifié leurs distributions. L'état de voisinage de la classe Urbaine montre que les Prairies et luzernes sont les voisins les plus fréquents de l'Urbain. Inversement, l'Urbain est un voisin fréquent des Prairies et luzernes, mais les voisins les plus fréquents (13%) sont des successions du type S(B,T,C) et plus particulièrement la rotation biennale : Blé-Tournesol qui représente, a elle seule, 5% du voisinage des Prairies et luzernes dans cette classe de patches.

S(B,T,C)		État de voisinage		S(U)		État de voisinage	
T-B-C-B	0.08	P-P-P-P	0.02	U-U-U-U	0.60	P-P-P-P	0.03
C-B-T-B	0.07	B-B-C-B	0.01	?-U-U-U	0.01	?-?-?-?	0.03
B-T-B-C	0.07	B-B-T-B	0.01	?-?-U-U	0.01	T-B-C-B	0.01
B-C-B-T	0.07	B-C-B-B	0.01	?-?-?-U	0.01	B-T-B-C	0.01
T-B-T-B	0.04	U-U-U-U	0.01	?-?-?-?	0.00	C-B-T-B	0.01
C-B-C-B	0.04	F-F-F-F	0.01	P-U-U-U	0.00	T-B-T-B	0.01
B-C-B-C	0.04	B-T-B-C	0.01	U-U-U-B	0.00	B-C-B-T	0.01
B-T-B-T	0.03	B-T-B-B	0.01	B-U-U-U	0.00	B-U-U-U	0.01
B-C-B-B	0.02	A-A-A-A	0.01	?-P-U-U	0.00	U-B-U-U	0.01
B-B-C-B	0.02	B-C-B-T	0.01	T-U-U-U	0.00	B-T-B-T	0.00

S(P)		État de voisinage	
P-P-P-P	0.20	T-B-T-B	0.03
Y-P-P-P	0.01	B-T-B-T	0.02
B-P-P-P	0.01	T-B-C-B	0.02
P-P-P-B	0.01	B-T-B-C	0.01
?-P-P-P	0.01	C-B-T-B	0.01
P-P-P-T	0.01	B-C-B-T	0.01
B-Y-P-P	0.01	B-T-B-B	0.01
T-P-P-P	0.00	U-U-U-U	0.01
P-P-T-B	0.00	B-B-T-B	0.01
T-B-P-P	0.00	C-B-C-B	0.01

TABLE 3 – Distributions des quadruplets dans les classes à l'issue d'une classification temporo-spatiale du corpus. ? désigne l'information manquante dans le corpus correspondant à des parcelles non enquêtées certaines années

5 Validation par étude des cliques

Pour mieux interpréter le contenu de l'état de réserve dans les HMM2 définissant les états du HHMM2 utilisé dans la classification spatiale, nous avons fouillé le graphe des voisinages des parcelles en étudiant l'évolution temporelle des couples (OCS d'une parcelle, OCS de la parcelle voisine) comme mentionné §2.4. Le parcours du graphe des voisinages est représenté par la courbe fractale réduite comme le font [1, 7] dans le cas d'un système de voisinage régulier. Une clique est un ensemble de parcelles deux à deux voisines, c'est à dire ayant une frontière commune. Une configuration des OCS à l'instant t d'une clique sera abusivement appelée "*clique temporelle*". Dans cet article, nous considérons des cliques simples faites de deux parcelles voisines représentées par leur centre de gravité. Des résultats préliminaires nous ont montré que la distribution des OCS est isotrope : la direction définie par les 2 centres de gravité ne porte pas d'information.

Deux sites successifs $(s_{l-1}, s_l),\quad 1 \leq l < L$ dans la courbe fractale réduite de longueur L définissent une clique. Sur chaque site s_l de la courbe fractale réduite, nous définissons un vecteur de traits o_l^t avec les OCS situées sur les cliques (s_{l-1}, s_l) :

$$o_l^t = \left(\left(s_{l-1}^t, s_l^t\right), \left(s_{l-1}^{t+1}, s_l^{t+1}\right)\right), \quad 0 \leq t < T-1, \quad 1 \leq l < L \tag{1}$$

s_l^t étant l'OCS au temps t et à l'index l dans la courbe fractale réduite. t représente le temps sur la période d'étude (12 ans) et l l'index spatial dans le parcours du plan.

Au temps t, un paysage est représenté par une suite de $L-1$ cliques temporelles se chevauchant. Nous considérons aussi $T-1$ représentations de paysage pour couvrir les T années de la période d'étude.

Les sites successifs sur la courbe fractale réduite résidant dans la même parcelle ne nous intéressent pas dans l'étude des voisinages entre parcelles. Les vecteurs de traits satisfaisant les propriétés $(s_{l-1}^t = s_l^t)$ et $(s_{l-1}^{t+1} = s_l^{t+1})$ seront supprimés dans les distributions.

Le vecteur de traits o_l^t est la réalisation de 4 variables aléatoires S_t, N_t, S_{t+1} et N_{t+1} qui définissent le processus stochastique observable. Ce vecteur se compare aux quadruplets temporels définis dans le §3.1 où les 4 OCS représentaient les 4 occupations successives d'une parcelle. A présent, ce vecteur représente 2 OCS successives sur 2 parcelles voisines.

La modélisation stochastique à l'aide d'un HMM2 permet d'extraire des patches caractérisés par des distributions de *cliques temporelles* (cf. Tab. 4).

– L'analyse des colonnes S_t et S_{t+1} montre les dépendances temporelles entre OCS sur un site alors que l'analyse des colonnes N_t et N_{t+1} montre les mêmes dépendances au niveau des parcelles voisines qui suivent donc la même logique de mise en valeur ;
– de même, l'analyse des colonnes S_t et N_t montre le degré d'attraction entre OCS ;
– enfin, l'analyse globale permet de quantifier l'attraction des successions d'OCS.

La table 4-a montre que les prairies sont les voisins "préférés" des zones urbaines comme le montrait déjà le traitement décrit au §4. Ce résultat nous permet de penser que l'état de réserve de chaque HMM2 (cf. §4) capture les successions à la lisière des patches. Le tableau montre aussi que ces 2 OCS sont stables dans le temps.

La table 4-b montre une nouvelle régularité que l'approche temporo-spatiale n'avait pas révélé. Les patches contenant les rotations incluant le tournesol (T), le blé (B) et le colza (C) dans des rotations bi ou quadri-annuelles sont constitués de parcelles dans lesquelles l'OCS voisine sera l'OCS suivante. Ceci montre que l'organisation dans le temps de la mosaïque agricole implique une organisation dans l'espace. Ceci nous conduit à envisager la description de la mosaïque en terme de quartiers culturaux formés de parcelles dans lesquelles les mêmes rotations sont pratiquées afin de réduire la variabilité temporelle et spatiale de cette mosaïque.

(a)

Freq.	S_t	N_t	S_{t+1}	N_{t+1}
0.014	U	P	U	P
0.014	P	U	P	U
0.007	T	U	B	U
0.007	U	T	U	B
0.007	U	B	U	C

(b)

Freq.	S_t	N_t	S_{t+1}	N_{t+1}
0.014	B	C	C	B
0.014	C	B	B	C
0.011	T	C	B	B
0.011	C	T	B	B
0.010	B	B	C	T
0.010	B	B	T	C
0.010	T	B	B	C
0.009	B	T	C	B
0.009	C	B	B	T
0.009	B	C	T	B

TABLE 4 – Distribution des cliques temporelles dans les patches urbains (a) et dans les patches contenant les successions à base de tournesol, blé et colza (b)

6 Discussion et conclusions

Dans cet article, nous avons développé deux approches visant chacune à décrire les relations de voisinages entre successions d'OCS à l'échelle du paysage agricole. Nous utilisons des modèles de Markov cachés pour modéliser la variabilité temporelle et spatiale du territoire agricole. Une première approche qualifiée de temporo-spatiale recherche des classes temporelles de successions d'OCS qui sont par la suite localisées après ajout d'un état de voisinage qui capture les successions à la lisière des patches ; la deuxième approche consiste à analyser les dépendances spatio-temporelles des successions d'occupation du sol en vue de valider et d'enrichir les connaissances extraites par la première approche.

Les modèles de classification mis au point dans ce travail visent à améliorer l'état des connaissances sur l'organisation territoriale de l'activité agricole. Ces modèles permettent de simplifier la complexité apparente de l'organisation et de l'évolution du paysage agricole à travers l'identification de zones homogènes en termes de successions d'occupation du sol mais aussi à travers l'identification de quartiers culturaux au sein desquels les occupations des parcelles voisines manifestent de forts degrés d'attraction. Chaque patche est décrit par deux distributions de successions : l'une décrivant son homogeneité spatiale, l'autre l'influence des patches voisins. Ceci nous permet de localiser la lisière du patche et permet une fouille plus détaillée, notamment une recherche de successions singulières ou innovantes.

A terme, ces techniques de classification développées et implémentées dans la boîte à outils ARPENTAGE sont à même de contribuer à améliorer les modèles de simulation des paysages visant à évaluer des risques divers dans des territoires agricoles à enjeux environnementaux comme les risques liés à la pollution de l'eau, à la diffusion des transgènes ou à la préservation d'espèces animales.

Remerciements

Nous remercions la région Lorraine, l'ANR BiodivAgrim, et l'API Ecoger pour leurs financements.

Références

[1] B. Benmiloud and W. Pieczynski. Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images. *Traitement du signal*, 12(5) :433 – 454, 1995.

[2] E.G. Lazrak, J.-F. Mari, and M. Benoît. Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology*, Sept. 2009. http ://hal.inria.fr/inria-00419952/en/.

[3] F. Le Ber, M. Benoit, C. Schott, J.-F. Mari, and C. Mignolet. Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software. *Ecological Modelling*, 191(1) :170 – 185, Jan 2006. http ://hal.archives-ouvertes.fr/hal-00017169/fr/.

[4] J.-f. Mari and C. Largouet. *Raisonnements sur l'espace et le temps : des modèles aux applications*, chapter 9 et 12, pages 249 – 271 et 305 – 316. ISBN : 2-7462-1455-5. Lavoisier, F. Le Ber edition, mars 2007. http ://hal.archives-ouvertes.fr/hal-00126022/fr/.

[5] J.-F. Mari and F. Le Ber. Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. *Soft Computing*, 10(5) :406 – 414, March 2006. http ://hal.inria.fr/inria-00000197.

[6] H. Memin and P. Perez. *Problèmes inverses en imagerie et vision*, chapter Estimation du mouvement, pages 205 – 267. Lavoisier, A. Mohammad-Djafari edition, 2009.

[7] W. Pieczynski. Markov models in image processing. *Traitement du signal*, 20(3) :255–278, 2003.

Chapitre 6

Articulation des régularités stochastiques avec les règles de décision des agriculteurs pour la modélisation des dynamiques de l'OTAA

Dans ce chapitre nous présentons des résultats de modélisation de la dynamique de l'OTAA à travers le développement d'une méthode combinant deux approches complémentaires : la modélisation des règles de décisions d'agriculteurs identifiées par enquêtes d'une part, et la modélisation de régularités stochastiques sur les proximités des OCS d'autre part. Ce chapitre comporte deux articles :

Le premier article (LAZRAK et al., 2011), destiné pour un public d'informaticiens, explique la méthode proposée en développant la méthode de fouille de données sans développer les aspects agronomiques.

Le deuxième article (SCHALLER et al., 2011) développe la méthode proposée du point de vue de l'agronome des territoires, en mettant l'accent sur la complémentarité entre les deux approches ainsi que sur les limites et perspectives d'amélioration de cette méthode.

Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol

El Ghali Lazrak* Noémie Schaller**
Jean-François Mari***

* Inra, UR 055 SAD ASTER, domaine du Joly, F-88500 Mirecourt
lazrak@mirecourt.inra.fr
** Inra/AgroParisTech, UMR 1048 SAD-APT, F-78850 Thiverval-Grignon
noemie.schaller@grignon.inra.fr
*** Loria, UMR CNRS 7503 et INRIA-Grand Est, F-54506 Vandœuvre-lès-Nancy
jfmari@loria.fr,
http://www.loria.fr/

Résumé. Nous modélisons la dynamique d'organisation spatiale et temporelle des paysages agricoles en articulant les échelles de l'exploitation agricole et du paysage. Nous développons une approche combinant deux méthodes : la modélisation des règles de décision d'agriculteurs obtenues par enquêtes d'une part et, d'autre part, la modélisation de régularités stochastiques sur les proximités des occupations du sol.

1 Introduction

Le paysage agricole peut être perçu comme un assemblage de polygones de tailles différentes – les parcelles – où chaque parcelle porte une occupation du sol (OCS). A l'échelle de l'exploitation agricole, la façon dont chaque agriculteur organise son territoire est un processus à la fois temporel et spatial qui modèle le paysage dans son ensemble. D'une façon symétrique, les changements temporels dans la mosaïque paysagère rendent compte de décisions des différents agriculteurs qui, sans concertation systématique mais de manière souvent convergente, mettent en valeur un territoire agricole en répondant à un ensemble de contraintes et opportunités. Nous présentons une méthode de fouille de données à l'aide de modèles stochastiques pour représenter la mosaïque agricole et comprendre son évolution temporelle et spatiale. Cette fouille s'appuie sur des enquêtes en exploitations agricoles qui alimentent les interactions entre l'analyste et les experts du domaine d'étude.

Pour analyser les dépendances temporelles et spatiales entre OCS, nous nous appuyons sur 2 hypothèses :

hypothèse de champ de Markov : l'OCS d'une parcelle dépend de l'OCS des parcelles voisines ;

hypothèse de chaîne de Markov : l'OCS d'une parcelle une année donnée dépend des OCS trouvées sur cette parcelle les années précédentes.

La fouille est dirigée par des experts agronomes qui, dans un premier temps à l'aide d'enquêtes effectuées dans les exploitations agricoles, retrouvent les traces dans la mosaïque paysagère des décisions prises à l'échelle de l'exploitation. Dans un second temps, une approche ascendante constate des régularités stochastiques dans la mosaïque et tente de les généraliser pour extraire des règles de décision qui n'avaient pas été préalablement énoncées. Cet article est structuré de la façon suivante : nous présentons dans une première partie les données de la fouille constituées d'enquêtes en exploitations agricoles et par un relevé systématique des occupations des parcelles agricoles. Dans une deuxième partie, nous présentons les méthodes mise en œuvre pour nettoyer ces données et réduire la dimension de l'espace de représentation. Nous présentons alors la mesure utilisée pour capturer la dynamique des voisinages entre cultures associées à des parcelles voisines. Les résultats, tirés d'un cas d'étude situé dans la plaine de Niort, sont donnés dans la section 4. Enfin, nous discutons de l'intérêt de cette approche hybride qui mêle enquêtes et modèles stochastiques pour fouiller des territoires agricoles et comprendre comment ceux-ci rendent compte des décisions prises, à une autre échelle, dans les exploitations agricoles.

2 Présentation des données

2.1 Construction d'un corpus d'OCS

Le paysage agricole étudié s'étend sur 350 km^2 dans la Plaine de Niort. Depuis plus de 12 ans, la localisation et les occupations des parcelles sont renseignées grâce à des relevés de terrain annuels. Pendant cette période d'étude, le territoire d'étude s'est étendu. Les parcelles nouvellement enquêtées – principalement proches des prairies – ont été étiquetées *indéterminé* les premières années avant de l'être par leurs véritables OCS. Ces relevés d'OCS annuels sont stockés dans un système d'information géographique sous format vectoriel et constituent une couche d'informations temporelles et spatiales.

Les frontières des parcelles changent chaque année en fonction des choix des agriculteurs (cf. Fig. 1 et 2). Pour tenir compte de ce changement, les enquêteurs définissent l'ensemble des micro-parcelles comme étant constituées de l'union de toutes les frontières de parcelles pendant la période d'étude. Il y a environ 20000 micro-parcelles dans le territoire étudié. Tous les points d'une micro-parcelles n'ont hébergé qu'une succession de cultures pendant la période d'étude. Dans la mosaïque parcellaire, le système de voisinage est irrégulier. Une parcelle a un nombre quelconque de parcelles avec lesquelles elle partage une frontière commune. Afin, d'éviter la complexité due à l'irrégularité du système de voisinage et à sa variabilité temporelle (cf. Fig. 2), nous avons choisi – dans une premier temps – de rastériser cette couche d'information vectorielle avec une grille de points régulièrement espacés (10m x 10m) dans les quatre directions cardinales. Le corpus résultant est une matrice où les colonnes représentent les OCS année par année et les lignes représentent les différents points d'échantillonnage localisés. Le corpus compte au total 47 modalités d'OCS que nous avons regroupées, dans un travail antérieur en 11 OCS (Lazrak et al., 2010) suivant une démarche tenant compte des fréquences des OCS et de la similitude des conduites culturales. Dans le présent travail, nous nous intéressons aux prairies en tant que voisins du tournesol et du maïs. Nous avons modifié le regroupement en individualisant les prairies et en classant l'orge d'hiver – non central pour cette étude – avec le blé afin de maintenir le même nombre des modalités (Tab. 1).

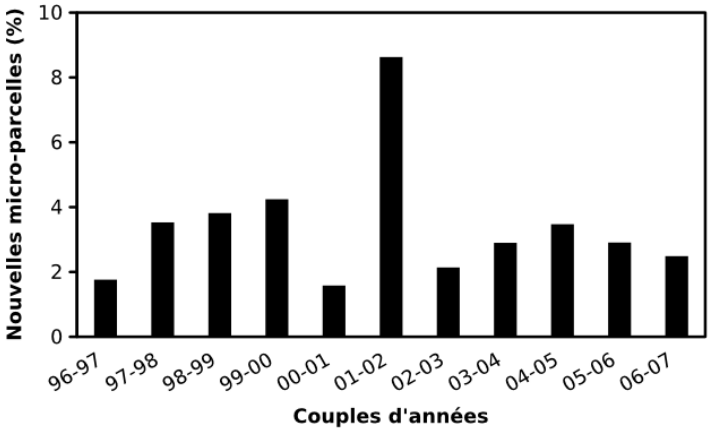


FIG. 1 – *Dynamiques inter-annuelles des frontières parcellaires dans la zone d'étude. Ces dynamiques sont exprimées en nombre de micro-parcelles nouvellement créées par rapport au nombre de parcelles de l'année précédente. Entre 2001 et 2002 plus de 8% des parcelles ont été redécoupées.*

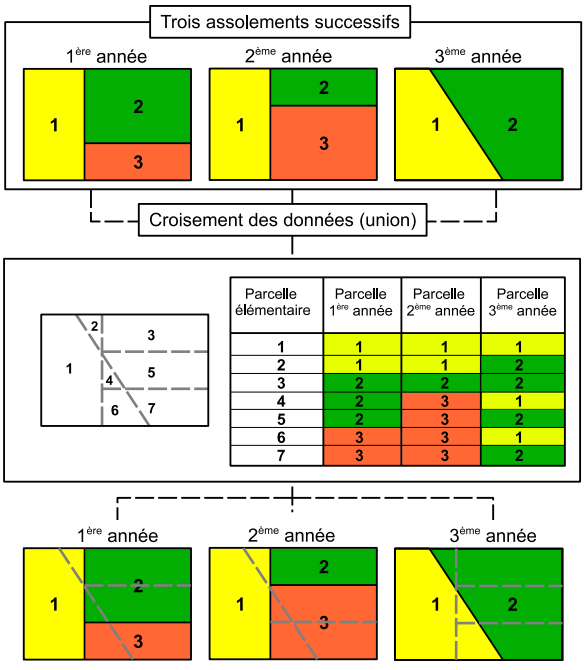


FIG. 2 – *Exemple d'évolution des limites de parcelles pendant trois années successives. L'union spatiale des frontières des parcelles pendant cette période aboutit à la définition de sept micro-parcelles.*

OCS initiales	Fréq. cumulée
Blé (B)	0.337
Tournesol (T)	0.476
Colza (C)	0.600
Urbain (U)	0.696
Prairies et Luzernes(P)	0.774
Maïs (M)	0.850
Forêts et friches (F)	0.884
Orge d'hiver (O)	0.918
Ray-grass (R)	0.942
Pois (S)	0.964
Autres (A)	1.000

OCS dans les enquêtes	Fréq. cumulée
Blé (B)	0.372
Tournesol (T)	0.511
Colza (C)	0.635
Urbain (U)	0.730
Maïs (M)	0.806
Prairies (P)	0.861
Forêts et friches (F)	0.896
Luzernes (L)	0.922
Ray-grass (R)	0.946
Pois (S)	0.968
Autres (A)	1.000

TAB. 1 – *Les OCS du paysage et leurs fréquences moyennes sur la période d'étude. Les enquêtes en exploitations ont nécessité de revoir les regroupements. A gauche le regroupement selon Lazrak et al. (2010). A droite le regroupement revu.*

2.2 Les enquêtes en exploitations agricoles

Afin d'analyser les logiques individuelles des agriculteurs, nous avons combiné deux modélisations conceptuelles : le "modèle pour l'action" d'une part (Sebillotte et Soler, 1990) et le modèle d'utilisation des ressources dans l'exploitation d'autre part (Aubry et al., 1998). A partir de ces modèles, nous avons construit un cadre conceptuel générique pour modéliser les décisions des agriculteurs à travers les variables de décision, les déterminants et les règles de décision (Schaller et al., 2010).

Les variables de décision permettent de décrire le contenu de la décision et donner une réponse à la question : "En quoi consiste la décision ?"

Les déterminants sont tous les éléments qui influencent les variables de décision. Ils peuvent être de différentes natures : quantitatifs ou qualitatifs, internes (par exemple les ressources de l'exploitation agricole) ou externes à l'exploitation agricole (par exemple les conditions du marché, le climat, ...).

Les règles de décision sont les règles qu'un agriculteur définit et suit, en fonction des déterminants, pour faire son choix et donner une valeur à chacune des variables de décision.

Pour une exploitation, les variables de décision relatives à l'allocation des cultures aux parcelles sont : (i) la zone cultivable de la culture définie par l'ensemble des parcelles adaptées à cette culture, (ii) la taille de la sole définie comme la surface totale d'une culture une année donnée sur l'exploitation, (iii) le délai de retour défini comme le temps minimum à attendre avant de replanter la même culture sur la même parcelle et (iv) les successions de cultures acceptables (Maxime et al., 1995; Navarrete et Le Bail, 2007; Merot et al., 2008).

Entre 2006 et 2010, nous avons réalisé 67 enquêtes parmi les 185 exploitations ayant toute leur surface dans la zone d'étude. Les enquêtes visaient à comprendre le fonctionnement global de l'exploitation agricole. Nous avons distingué quatre objectifs spécifiques, qui ont été atteints grâce à quatre sessions successives d'enquêtes :

– 22 enquêtes en 2006 et 19 enquêtes en 2007 ont porté respectivement sur les stratégies des éleveurs et des agriculteurs pour faire face aux sécheresses estivales et aux interdictions d'irrigation (Havet et al., 2010) ;

– 12 enquêtes en 2009 ont porté sur les décisions des agriculteurs relatives à l'allocation des cultures dans les parcelles et au découpage des parcelles ;
– 14 enquêtes en 2010 ont porté sur l'évolution dans le temps des assolements annuels choisis par l'exploitant.

Les enquêtes étaient semi-structurées pour encourager l'agriculteur à expliciter les raisons de ses choix et leur évolution au fil du temps, notamment la façon d'allouer les successions de cultures dans les parcelles.

3 Méthodes

3.1 Choix de l'observation élémentaire

Le choix de l'observation élémentaire permet de définir les modalités d'un pixel de l'image représentant la mosaïque agricole. Plusieurs observations élémentaires sont envisageables :

1. l'OCS en un point d'une parcelle représentant son occupation ;
2. la succession d'OCS en un point d'une parcelle sur deux ou plusieurs années successives. L'observation élémentaire est un n-uplet d'OCS se chevauchant temporellement ;
3. l'OCS en un point d'une parcelle augmenté de ses 4 voisins du premier ordre : Nord (No), Sud (So), Est (Es), Ouest (We). Les observations élémentaires sont des quintuplés d'OCS se chevauchant spatialement ;
4. le couple (OCS, OCS d'une parcelle voisine). Les observations élémentaires sont les configurations des cliques – deux sites voisins – se chevauchant spatialement.

La première observation est pratique pour calculer l'assolement moyen et l'évolution temporelle de celui-ci (Mari et Le Ber, 2006; Mignolet et al., 2007).

La seconde observation est utile pour retrouver les successions dominantes selon la méthode de fouille développée par (Le Ber et al., 2006; Lazrak et al., 2010).

La troisième observation permet de calculer l'information mutuelle spécifique entre OCS voisines (cf. 6), de tester l'isotropie du paysage et de déterminer la résolution spatiale optimale en fonction de la diversité des observations (Figure 3).

Enfin, l'utilisation de couples (OCS, OCS d'une parcelle voisine) permet de fouiller les voisinages entre OCS et leurs évolutions d'une façon efficace lorsque le milieu est isotrope. Cette information élémentaire permet de réduire significativement le nombre d'observations différentes, et de diminuer l'encombrement mémoire nécessaire pour représenter les distributions des observations dans les modèles stochastiques de fouille élaborés.

4 Modélisation stochastique à l'aide de HMM2

Afin d'éviter le biais introduit par la modalité *indeterminé* au voisinage des prairies pendant les premières années de l'étude, nous effectuons une segmentation des séquences d'OCS par un HMM2 afin de d'isoler ce segment temporel d'*indéterminé*. Nous effectuons un alignement élastique de la séquence des 12 OCS avec un HMM2 linéaire chargé de capturer les OCS *indéterminé* dans ses premiers états. Nous utilisons des modèles de Markov cachés du second ordre HMM2 (Mari et Le Ber, 2006) pour représenter la dynamique temporelle des voisinages

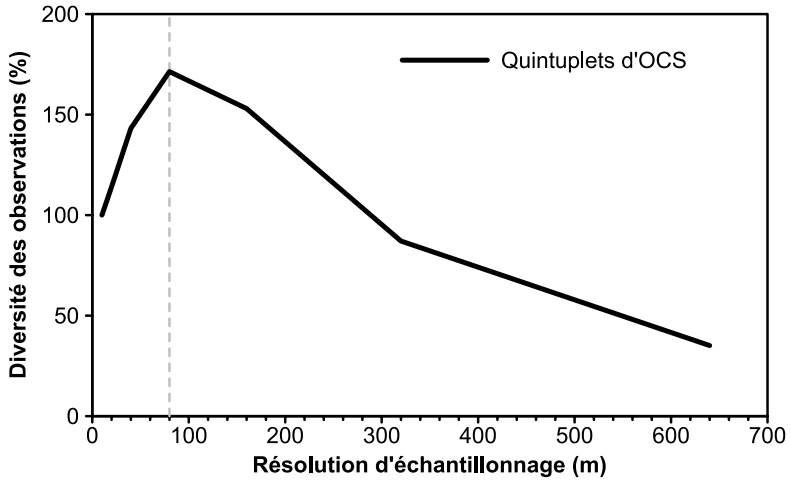


FIG. 3 – *Nombre de quintuplés (la configuration d'un site augmenté de ses 4 voisins) suivant la résolution spatiale. Le nombre de quintuplés à 10 m est considéré comme référence (100%). La courbe montre les mêmes propriétés qu'en analyse de textures d'images numériques. 80 m est la résolution spatiale donnant la plus grande diversité de voisinages et sera retenue dans la suite de cette étude*

représentés par des quintuplés d'OCS. Chaque année t, pendant une période de 12 ans, un site S_t et ses 4 voisins (Nord, Sud, Est, Ouest) prennent 12 valeurs de quintuplés différentes représentées par les 5 variables aléatoires $(S_t, No_t, So_t, Es_t, We_t, \quad t = 1, 12)$. Nous modélisons cette suite à l'aide d'un HMM2 linéaire estimé sur tous les sites du territoire étudié. La modélisation reprend les principes donnés dans (Le Ber et al., 2006; Mari et Le Ber, 2006) et cherche à segmenter la période en autant de classes que d'états. Nous cherchons des segments temporels non chevauchants pendant lesquels la distribution des quintuplés est stationnaire. L'estimation se fait selon le maximum de vraisemblance en utilisant l'algorithme forward-backward. La figure 4 montre les différentes associations des années avec les états du HMM2 suivant le nombre d'états. Par exemple, cette figure montre qu'un modèle de 7 états permet une association biunivoque entre les états et les segments et d'associer chaque état à un seul segment temporel d'une durée moyenne de 2 ans. La proportion d'*indéterminé* est maximale dans les premiers états. Les segments temporels associées sont ignorés.

5 Voisinages et cliques

Pour représenter la relation de voisinage entre sites, nous estimons la probabilité conditionnelle $P(V/S)$ représentant la probabilité d'avoir l'OCS x sur le site voisin – $(V = x)$ – sachant le site actuel occupé par l'OCS y : $(S = y)$. Ces probabilités sont estimées à partir des lois marginales des distributions des quintuplés. Si les distributions jointes $P(S, V)$ sont les mêmes quelle que soit la direction du voisinage – No, So, Es, We – la mosaïque agricole est

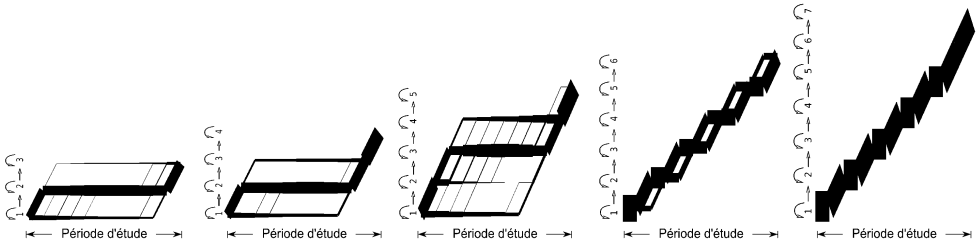


FIG. 4 – *Recherche d'une segmentation satisfaisante de la période d'étude avec des HMM2 linéaires ayant un nombre croissant d'états. L'épaisseur des traits est proportionnelle à la probabilité a posteriori des états du HMM2. La segmentation recherchée comporte au moins 2 périodes de plusieurs années ne se recouvrant pas et ne contenant pas l'OCS* indéterminé. *Dans un HMM2 de 6 états, les états 2 et 5 identifient deux périodes disjointes*

dite isotrope. La distance entre deux distributions est calculée à l'aide de la divergence (Tou et Gonzales, 1974)

$$div(f,g) = \frac{1}{2}\sum_{x}(f(x) - g(x)) \log \frac{f(x)}{g(x)} \quad (1)$$

quand f et g sont deux distributions discrètes sur le même espace décrit par x.

A la dernière itération de l'algorithme *Forward-backward*, les comptes des quintuplés sont calculés sur chaque état et permettent le calcul des comptes des cliques Nord (S, No), Sud (S, So), Est (S, Es) et Ouest (S, We). A partir de ces comptes, on estime les lois marginales $P(S)$ et $P(V,S)$.

Les seules cliques que nous considérons sont constituées de deux sites voisins : – soit horizontalement, soit verticalement – de configurations différentes : on ignore les cliques "plein champ" dont la configuration est faite de deux OCS identiques. Cela revient à n'échantillonner le territoire que le long des frontières des parcelles occupées par des OCS différentes. Une fois l'isotropie du paysage démontrée, nous considérons que l'orientation ne porte plus d'information et nous utilisons la clique d'OCS comme observation élémentaire pour fouiller les relations de voisinage entre cultures et leur évolution.

6 Information mutuelle spécifique

La probabilité du voisinage $P(V/S)$ n'est pas une bonne mesure pour évaluer la co-localisation de deux OCS car elle dépend des probabilités marginales des OCS. Nous utilisons l'information mutuelle spécifique (PMI comme Pointwise Mutual Information) (Novovičová et al., 2004) définie de la façon suivante :

$$PMI(x,y) = \log(\frac{P(V=x/S=y)}{P(V=x)}) = \log(\frac{P(V=x,S=y)}{P(V=x)\times P(S=y)}) \quad (2)$$

Cette quantité représente l'information apportée par la connaissance d'une variable sur l'autre. Une valeur positive signifie que le couple d'OCS (x,y) est co-localisé : les OCS x et y

s'attirent. L'expert du domaine (l'agronome) peut dans ce cas rechercher les règles de décision des agriculteurs qui expliquent cette co-localisation. Une valeur nulle signifie que les variables V et S sont indépendantes. Une valeur négative signifie que les OCS se repoussent, l'agronome peut dans ce cas expliquer ou rechercher par enquêtes auprès des agriculteurs la (les) raison(s) d'éviter de mettre ces OCS côte à côte.

La PMI est une mesure qui se rencontre dans d'autres domaines : dans l'analyse du texte écrit (Schneider, 2005) pour la recherche des couples de mots co-localisés, et aussi en analyse d'images quand il est question d'étude des voisinages ou des textures.

7 Résultats

7.1 Segmentation temporelle

En fonction des résultats donnés Fig. 3, nous choisissons une résolution de 80 m qui donne la plus grande diversité de voisinages, représentées par environ 40000 quintuplés différents d'OCS. A cette résolution, l'ensemble des sites est utilisé pour l'apprentissage de différents HMM2, comme le montre la figure 4. Sur chacun des états, nous calculons les lois marginales $P(V, S)$ dans chaque direction. La matrice des divergences obtenue est nulle (de l'ordre de 10^-2) sur chacun des états et confirme l'hypothèse d'isotropie de la mosaïque agricole. Par la suite, nous estimons un HMM2 à l'aide de cliques d'OCS comme observations élémentaires sans tenir compte de leur orientation. Nous calculons la PMI à partir des comptes des cliques sur les états sélectionnés.

Le HMM2 linéaire retenu comporte 6 états et permet de définir 6 périodes différentes parmi lesquelles les périodes correspondants aux états 2 et 5 ne se recouvrent pas. L'état 2 correspond à la période 1998 à 2000, et l'état 5 correspond à la période de 2004 à 2006. Nous avons choisi ces états pour comparer les relations de voisinage du tournesol et du maïs sur deux périodes distinctes encadrant une période de sécheresse qui a influencé le raisonnement des agriculteurs.

7.2 Impact des décisions prises au niveau des exploitations sur le paysage agricole

Dans cette étude, le cadre conceptuel des enquêtes (variable, déterminant, règle) et le cadre Markovien sont liés. Le cadre formel des enquêtes a fait apparaître des règles de décision chez les agriculteurs dont l'impact sur le paysage est évalué par la modélisation Markovienne.

- Les variables "couples précédent/suivant" et "délai de retour" permettent d'exploiter la dimension temporelle des décisions, et donc des régularités en termes de successions de cultures ;
- la variable "zone cultivable" permet d'explorer la dimension spatiale des décisions, et donc des régularités en termes de voisinage de cultures ;
- la variable "taille de sole", permet, le cas échéant, d'explorer l'évolution au cours du temps des surfaces des catégories de cultures.

Dans un premier temps, les méthodes d'enquêtes ont révélé une règle de décision commune entre exploitations agricoles concernant la localisation de la culture de tournesol : les agriculteurs évitent de cultiver le tournesol à proximité des forêts et des bosquets en raison des dégâts plus fréquents causés par les ravageurs (lapins, corbeaux). La figure 5 montre que le voisinage

entre tournesol et forêts (T-F) est moins fréquent au cours de la période 2004-2006 (état 5 du HMM2) qu'en 1998-200 (état 3 du HMM2) contrairement aux voisinages avec les cultures de vente telles que blé (T-B) ou colza (T-C).

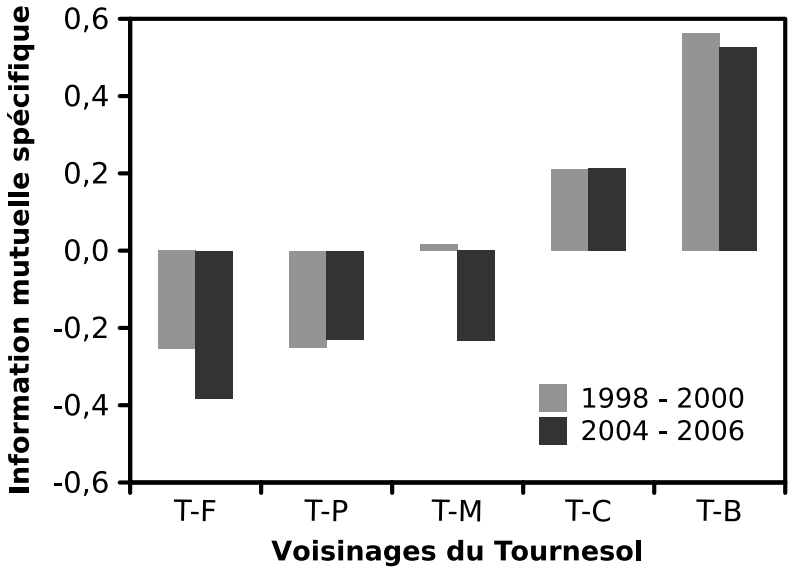


FIG. 5 – *Évolution de l'information mutuelle spécifique entre le tournesol (T), les prairies (P), le colza (C) et le blé (B). Plus la PMI est élevée, plus les cultures sont fréquemment voisines*

7.3 Apparition de nouvelles règles de décision par fouille des évolutions

La modélisation stochastique a également permis d'identifier une régularité d'évolution du voisinage entre maïs et prairies. Ces deux OCS ont tendance à être de plus en plus fréquemment voisines tandis que maïs et colza ou tournesol sont de moins en moins fréquemment voisins (Figure 6).

Cette régularité semble cohérente avec une règle de décision commune identifiée concernant le maïs : les agriculteurs réduisent les surfaces en maïs en raison des risques de sécheresse estivale et le concentrent dans les terrains les plus humides, fréquemment à proximité des prairies. Les éleveurs étendent même la surface des prairies pour sécuriser la production de fourrage dans le cas où la production de maïs serait insuffisante, d'où la co-localisation de ces deux OCS.

Ainsi, les résultats obtenus par enquêtes et modélisation stochastique apparaissent cohérents, ce qui suggère une bonne complémentarité entre les deux méthodes pour modéliser les dynamiques d'organisation spatiale des paysages. Les règles de décision identifiées à l'échelle de l'exploitation peuvent être évaluées à l'échelle du paysage, tandis que les régularités stochastiques du paysage pourraient être en partie expliquées par des règles de décision d'agriculteurs.

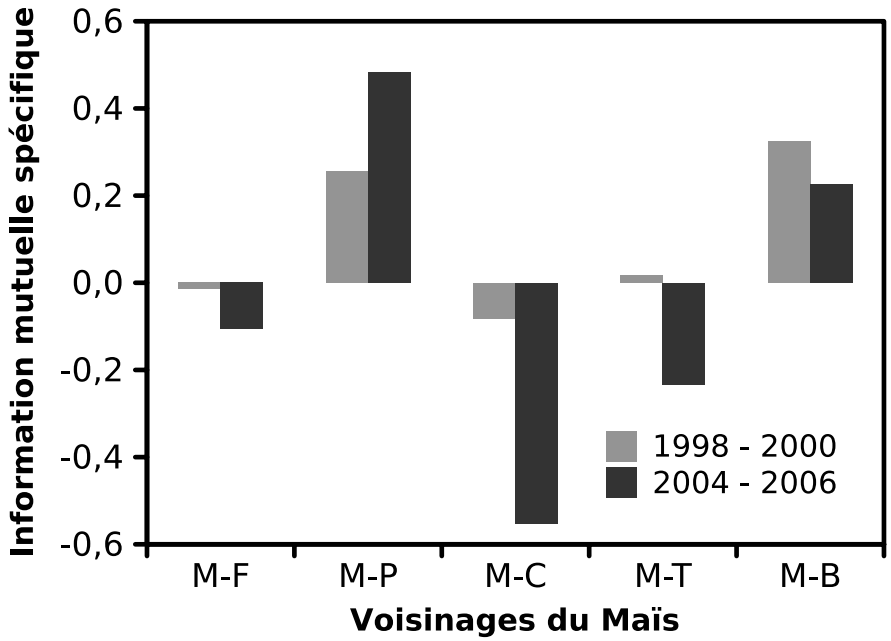


FIG. 6 – *Évolution de l'information mutuelle spécifique du maïs avec les autres OCS. Celui-ci s'éloigne des forêts qui abritent les ravageurs. On remarque aussi l'évolution vers une co-localisation avec les prairies (P) et sa disparition des terres "à colza" (C)*

8 Conclusions

Nous avons présenté une méthode de fouille de données complexes pour identifier et modéliser les règles de décision d'agriculteurs à l'échelle de l'exploitation agricole concernant l'assolement et pour retrouver leurs impacts dans le paysage sous forme de régularités stochastiques.

Les données de la fouille provenaient de deux sources différentes : une source est constituée d'enquêtes effectuées sur un échantillon d'exploitations agricoles portant sur des variables, des déterminants et des règles aboutissant aux décisions d'assolement et l'autre source est constituée de relevés exhaustifs d'occupation du territoire qui rendent compte de la variabilité dans le temps et l'espace des parcelles agricoles et leurs occupations. L'information traitée se situait à deux niveaux d'échelle.

Pour contrer le fait que le territoire d'étude n'a pas été enquêté uniformément dans le temps, nous avons segmenté la période d'étude en plusieurs sous périodes par un HMM2 qui permet d'isoler les OCS "indéterminées" et de déterminer deux périodes d'études non chevauchantes.

Après avoir montré que le territoire est isotrope vis à vis des OCS, le calcul de la PMI sur les configurations des cliques a permis de montrer les tendances au rapprochement (phénomène de co-localisation) ou à l'éloignement entre OCS que des enquêtes dans les exploitations peuvent expliquer.

Les dynamiques d'organisation spatiale des paysages agricoles impactent de nombreux processus environnementaux. Modéliser les paysages agricoles est donc une étape clé pour pouvoir décrire et comprendre ces dynamiques d'organisation spatiale des paysages, ainsi que leurs conséquences environnementales. Une perspective importante de ce travail est ainsi d'utiliser les règles de décision d'agriculteurs et les régularités stochastiques pour générer des paysages agricoles et tester des scénarios. A terme, cette perspective pourrait permettre aux gestionnaires des territoires agricoles d'agir sur les décisions des agriculteurs afin d'orienter favorablement les dynamiques d'organisation spatiale des paysages pour des questions environnementales locales.

Remerciements

Nous remercions le Centre d'étude biologique de Chizé (CEBC UPR 1934 CNRS), les régions Lorraine et Île de France, l'ANR BiodivAgrim et l'API Ecoger pour leurs supports.

Références

Aubry, C., A. Biarnes, F. Maxime, et F. Papy (1998). Modélisation de l'organisation technique de la production dans l'exploitation agricole : la constitution de système de culture. *Etud Rech Syst Agraires Dév* (31), 25–43.

Havet, A., P. Martin, M. Laurent, et B. Lelaure (2010). Adaptation des exploitations laitières aux incertitudes climatiques et aux nouvelles réglementations. le cas des productions bovines et caprines en plaine de niort. *Fourrages 202*, 145–151.

Lazrak, E., J.-F. Mari, et M. Benoît (2010). Landscape regularity modelling for environmental challenges in agriculture. *Landscape Ecology 25*(2), 169 – 183. http ://hal.inria.fr/inria-00419952/en/.

Le Ber, F., M. Benoît, C. Schott, J.-F. Mari, et C. Mignolet (2006). Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software. *Ecological Modelling 191*(1), 170 – 185. http ://hal.archives-ouvertes.fr/hal-00017169/fr/.

Mari, J.-F. et F. Le Ber (2006). Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. *Soft Computing 10*(5), 406 – 414. http ://hal.inria.fr/inria-00000197.

Maxime, F., J. Mollet, et F. Papy (1995). Aide au raisonnement de l'assolement en grande culture. *Cah Agri* (4), 351–362.

Merot, A., J. Bergez, A. Capillon, et J. Wery (2008). Analysing farming practices to develop a numerical, operational model of farmers' decision-making processes : An irrigated hay cropping system in France. *Agricultural Systems 98*(2), 108–118.

Mignolet, C., C. Schott, et M. Benoît (2007). Spatial dynamics of farming practices in the Seine basin : Methods for agronomic approaches on a regional scale. *Science of The Total Environment 375*(1–3), 13–32. http ://www.sciencedirect.com/science/article/B6V78-4N3P539-2/2/562034987911fb9545be7fda6dd914a8.

Navarrete, M. et M. Le Bail (2007). Saladplan : a model of the decision-making process in lettuce and endive cropping. *Agron Sust Dev 3*(27), 209–221.

Novovičová, J., A. Malik, et P. Pudil (2004). Feature selection using improved mutual information for text classification. In A. Fred, T. Caelli, R. P. W. Duin, A. Campilho, et D. d. Ridder (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition*, Volume 3138 of *Lecture Notes in Computer Science*, pp. 1010–1017. Springer Berlin / Heidelberg.

Schaller, N., C. Aubry, et P. Martin (2010). Modelling farmers' decisions of splitting agricultural plots at different time scales : a contribution for modelling landscape spatial configuration. Montpellier, France, pp. 879–880.

Schneider, K.-M. (2005). Weighted Average Pointwise Mutual Information for Feature Selection in Text Categorization. pp. 252–263.

Sebillotte, M. et Soler (1990). *Modélisation systémique et systèmes agraires*, Chapter Les processus de décision des agriculteurs : acquis et questions vives, pp. 93–102. INRA Paris.

Tou, J. T. et R. Gonzales (1974). *Pattern Recognition Principles*. Addison-Wesley.

Summary

We model the dynamics of space and temporal organization of agricultural landscapes by articulating the farm and landscape levels. We develop an approach combining two methods : the modeling of the decision rules of farmers acquired by on farm surveys and the stochastic modeling of neighborhood regularities.

Landscape Ecol (2012) 27:433–446
DOI 10.1007/s10980-011-9691-2

RESEARCH ARTICLE

Combining farmers' decision rules and landscape stochastic regularities for landscape modelling

Noémie Schaller · El Ghali Lazrak · Philippe Martin · Jean-François Mari · Christine Aubry · Marc Benoît

Received: 7 January 2011 / Accepted: 3 December 2011 / Published online: 18 December 2011

Abstract Landscape spatial organization (LSO) strongly impacts many environmental issues. Modelling agricultural landscapes and describing meaningful landscape patterns are thus regarded as key-issues for designing sustainable landscapes. Agricultural landscapes are mostly designed by farmers. Their decisions dealing with crop choices and crop allocation to land can be generic and result in landscape regularities, which determine LSO. This paper comes within the emerging discipline called "landscape agronomy", aiming at studying the organization of farming practices at the landscape scale. We here aim at articulating the farm and the landscape scales for landscape modelling. To do so, we develop an original approach consisting in the combination of two methods used separately so far: the identification of explicit farmer decision rules through on-farm surveys methods and the identification of landscape stochastic regularities through data-mining. We applied this approach to the Niort plain landscape in France. Results show that generic farmer decision rules dealing with sunflower or maize area and location within landscapes are consistent with spatiotemporal regularities identified at the landscape scale. It results in a segmentation of the landscape, based on both its spatial and temporal organization and partly explained by generic farmer decision rules. This consistency between results points out that the two modelling methods aid one another for land-use modelling at landscape scale and for understanding the driving forces of its spatial organization. Despite some remaining challenges, our study in landscape agronomy accounts for both spatial and temporal dimensions of crop allocation: it allows the drawing of new spatial patterns coherent with land-use dynamics at the landscape scale, which improves the links to the scale of ecological processes and therefore contributes to landscape ecology.

Keywords Land-use dynamic · On-farm survey · Conceptual model · Data mining · Crop succession · Crop allocation · Spatiotemporal analysis · Landscape agronomy · Landscape patterns

N. Schaller (✉) · P. Martin · C. Aubry
AgroParisTech, INRA, UMR 1048 SAD-APT, Bâtiment EGER, BP 01, 78 850 Thiverval-Grignon, France
e-mail: noemie.schaller@agroparistech.fr

E. G. Lazrak · M. Benoît
INRA, UR 055 SAD ASTER, 88500 Mirecourt, France

J.-F. Mari
LORIA, UMR CNRS 7503 INRIA-Grand-Est, B.P. 239, 54506 Vandœuvre-lès-Nancy, France

Introduction

Agriculture is the most important land use across Europe (Rounsevell et al. 2003) and its impacts on the environment are no longer to be demonstrated (Donald et al. 2001; Stoate et al. 2001). Depending on market

conditions, access to technology and public policies, subsequent intensification or abandonment can have contrasting effects from one region to another and even within regions (Stoate et al. 2009). Considering the landscape scale of agroecosystems is thus necessary to address many ecological processes.

Landscape spatial organization (LSO) strongly impacts many environmental issues. Agricultural landscapes in Europe are composed of a crop mosaic and of semi-natural areas (Burel and Baudry 2010). We focus here on the crop mosaic, directly influenced by agricultural practices and we define the LSO as the spatial structure and arrangement of the agricultural plots within the landscape. Several studies have shown that modifying the LSO can orientate environmental processes like biodiversity preservation (Benton et al. 2003; Joannon et al. 2008), soil erosion by water and tillage (van Oost et al. 2000), erosive runoff (Joannon et al. 2006), water pollution (Benoît et al. 1997; Beaujouan et al. 2001) and gene fluxes (Le Bail et al. 2010).

In order to understand the interactions between LSO and ecological processes, it is necessary to identify and describe meaningful landscape patterns (Turner 1990), both for scientists and planners. The identification and description of such landscape patterns could improve the understanding of environmental processes in relation with agricultural dynamics, which may facilitate the exploration of the future and political decision making (Lazrak et al. 2010a). Modelling agricultural landscapes is thus a key-step in exploring land-use dynamics and helping design sustainable and environmentally-friendly landscapes (Veldkamp and Lambin 2001; Gaucherel and Houet 2009).

Agricultural landscapes are primarily designed by farmer practices (Benoît 1990; Le Ber and Benoît 1998; Thenail et al. 2009). Farmer decisions dealing with crop choices and crop allocation to land at farm scale impact LSO (Thenail and Baudry 2004). As a consequence, there is a spatial mismatch between the landscape scale, where environmental processes occur and should be managed, and the farm scale, where landscape units are managed through farmer practices (Rindfuss et al. 2004; Cumming et al. 2006; Pelosi et al. 2010). This mismatch calls for a better articulation between the farm and the landscape scales.

As agronomists, we usually focus on farmer practices and soil-crops-climate interactions at the farm scale. Our research here focuses on the role of farmers in the landscape design process and is thus part of an emerging branch of agronomy called "landscape agronomy" (Benoît et al. 2007). This discipline focuses on the landscape scale and aims at studying the organization of farming practices on a small geographical scale (Mignolet et al. 2007). The scale generally ranges from 100 km^2 to a few thousand km^2 and is thus intermediate between the farm scale (0.1–10 km^2) and coarser scales (>100,000 km^2). Landscape agronomy has recently developed due to the increased attention given to localised environmental problems resulting from farming activity (Benoît et al. 2007). It combines concepts and methods from geographers and agronomists: multi-scale modelling approaches for land-use changes (Veldkamp and Fresco 1996; De Koning et al. 1999; Lambin et al. 2003) and analytical methods to describe the underlying reasoning of regional agricultural systems organization. It relies mostly on the spatialization of farming system classifications (Landais 1998; Leisz et al. 2005; Mignolet et al. 2007).

Farmer decisions dealing with crop choices, crop successions and crop allocation to land have already been modelled at farm scale (Maxime et al. 1995; Aubry et al. 1998a; Navarrete and Le Bail 2007; Mawois et al. 2011). Besides, some authors have shown that farmers organize their crops in the farming territory according to spatial constraints (Morlon and Benoît 1990; Morlon and Trouche 2005), especially on mixed crop-livestock farms (Brunschwig et al. 2006; Marie et al. 2009). It is now accepted that the way a farmer organizes his farming territory is related to his cropping plan and is both a time and a spatial process (Dury et al. 2011).

Even if taken at the individual level, some decisions can be generic (i.e. shared by a set of farmers) and may result in landscape spatiotemporal regularities. Such regularities have already been stochastically modelled at the landscape scale (Lazrak et al. 2010a). These authors consider that land-uses are heterogeneously distributed among different polygons (agricultural plots) across landscapes and these land-uses display dynamic patterns as a result of crop successions and other driving forces of land-use changes. These patterns can be modelled both in their spatial and temporal dimensions using a Markovian framework (Le Ber et al. 2006; Mignolet et al. 2007; Castellazzi et al. 2008; Lazrak et al. 2010a) or stochastic decision trees (Sorel et al. 2010).

In short, there are methods for identifying (i) farmer decisions at farm scale and (ii) regularities in spatial and

temporal patterns at landscape scale, but very few studies aim at articulating the farm and the landscape scales by linking the two methods (Pocewicz et al. 2008). Agent-based models have been widely used to simulate land-use changes as a result of human decisions, but these approaches do not focus on farmers as the main stakeholders in agricultural landscapes. Agent-based models hardly account for farmer technical management: they rather focus on the impact of economic factors (e.g. agricultural and development policies, land ownership) on farm structure or land cover changes (Freeman et al. 2009; Happe et al. 2009; Robinson and Brown 2009). Upscaling to the regional scale is even possible thanks to probabilistic approaches (Valbuena et al. 2010), but disregarding the diversity of farmer practices (e.g. crop successions).

Following these authors, we now hypothesize that farmers play the most crucial role in designing agricultural landscapes (Bacic et al. 2006) and that combining methods for identifying farmer decisions and landscape regularities could bridge the gap between crop patterns generated by farmers and agricultural landscape modelling. The aim of the paper is thus to model farmer decisions and landscape regularities, as well as the links between them, in order to improve the articulation of the farm and the landscape scales for regional land use modelling. To do so, we used a French case study, where researchers more widely aim at understanding (i) how farmers make their crop allocation choices at farm scale and the associated determinants (ii) what landscape regularities can be identified at landscape scale in relation to farmer decisions, and (iii) how the landscape is finally designed and spatially organized over time. Once we have presented each of the two modelling methods, we will show how they mutually benefit one another before discussing advances, future challenges and perspectives for landscape modelling and research.

Methods

Study area

We applied our approach to the case of the Niort plain landscape in France. This area is located in the South of Deux-Sèvres in the Poitou–Charentes region (46.2°N, 0.4°W). Its extent is about 350 km^2 (Lazrak et al. 2010a), the average plot area being about 4–5 ha.[1] This agricultural landscape comprises woods and villages (4% of the total area) and is mainly composed of cereals, maize, sunflower, rapeseed and of a minority of grasslands (15%). The number of mixed crop-livestock farms (cattle and goats) has indeed been reduced in favour of arable farms. The Niort plain is a European Natura 2000 area where LSO particularly matters for biodiversity preservation and both water quality and quantity issues.

Methods at farm scale

On-farm surveys and farmer decision rules

Farmer decisions are qualitative data: we modelled farmer decisions through a generic conceptual framework describing such decisions as (i) decisional variables (ii) determinants and (iii) decision rules. A decision rule refers to a decisional variable and is influenced by one or several determinants:

(i) Decisional variables (Aubry et al. 1998a, 2006; Navarrete and Le Bail 2007) describe the content of the decision and give an answer to the question: "what is the decision about?" For allocating crops to land at farm scale, decisional variables have been identified as: suitable cultivation area for each crop (all suitable plots for the considered crop), crop area (total area of a considered crop on the farming territory), crop return time (acceptable time to replant the same crop on the same plot) and preceding–following crop pairs (acceptable temporal crop sequences) (Maxime et al. 1995; Aubry et al. 1998a; Navarrete and Le Bail 2007).

(ii) Determinants are all elements influencing the decisional variables and the decision rules: they can be of different natures (quantitative or qualitative) and internal (e.g. farm resources) or external to the farm (e.g. market conditions, climate). For example, on-farm labour force and market prices can determine the choice of crop areas.

(iii) Decision rules (Sebillotte and Soler 1990; Aubry et al. 1998b; Merot et al. 2008) are the rules each farmer defines and follows, depending on the determinants, to make his choice for

[1] http://www.zaplainevaldesevre.fr/index.php.

each decisional variable. For example, concerning the decisional variable "crop area", a rule could be "if the price of this crop is lower than [a threshold], then I will decrease this crop area". These rules may be qualitative but their content can be compared between farmers for a given decisional variable and reveal that some farmers share the same rules. Such generic rules will be described in the results section.

In this study, we applied this general rule-based model to specifically analyse decisions dealing with crop choices and crop allocation to land. The necessary data for such a model requires specific on-farm surveys (Merot et al. 2008). Between 2006 and 2010, we carried out 67 surveys in the Niort plain landscape. We sampled 67 farms out of the 185 farms having the whole farming territory inside the study area. The sample was built in order to account for the diversity in farming systems and not to be representative of all farms. All on-farm surveys aimed at understanding the global functioning of the farm. In addition, we distinguished four specific goals, which were achieved through four successive sessions of on-farm surveys: 22 surveys in 2006 and 19 surveys in 2007 focused respectively on mixed crop-livestock and arable farmer strategies to cope with summer droughts and irrigation bans (Martin et al. 2009; Havet et al. 2010); 12 surveys in 2009 focused on farmer decisions dealing with crop spatial allocation and plot splitting (Schaller et al. 2010); 14 surveys in 2010 focused on the evolution over time of farmers' annual cropping plans. The 67 on-farm surveys were semi-structured to encourage farmers to specify the reasons for their choices and how these choices could evolve over time, especially regarding the way they allocated crops to land.

The on-farm understanding of such rotational principles and land allocation are now considered as the driving factors of landscape patterns (Thenail et al. 2009). Such patterns can be detectable by statistical methods (Castellazzi et al. 2007) and data mining methods (Mignolet et al. 2007; Lazrak et al. 2010a).

Methods at landscape scale

The land-use data-base

To identify landscape stochastic regularities, we used a data-base providing information about the land-use in the Niort plain. This data-base was built by the Chizé Centre for Biological Studies, based on biannual land-use surveys covering an extent of 350 km^2 and going back to 1996 (Lazrak et al. 2010a). The two surveys in April and June made it possible to account for both early-harvested and late-planted crops. Each year, surveyors distinguished 47 land-uses (42 agricultural, three urban and two forest land-uses) and updated the plot limits when necessary (Lazrak et al. 2010a). The land-use surveys resulted in a GIS geodatabase in vector format.

Theoretical background for modelling landscape regularities

In order to model temporal and spatial landscape regularities, we used a stochastic data-mining approach based on a Markovian framework. Stochastic modelling for data mining is a convenient way of building statistical and probabilistic models for capturing the spatiotemporal data variability that is not yet fully understood. This Markovian framework is based on two assumptions in spatial and temporal domains respectively: (i) the *Markov random field* (MRF) *assumption* assumes that the land-use of a given field depends only on the land-use of the neighbouring fields; (ii) the *Markov chain assumption* assumes that the land-use of a given field in a year depends only on the land-use of the recent previous years in the same field. We used second order Hidden Markov Models (HMM2) to approximate the Markov assumptions, assuming that the distribution observations (land-uses) in an area at time t—the cropping plan—depend on the cropping plan observed at time $t - 1$ or $t - 2$. HMM generalize Markov chains (Castellazzi et al. 2008) through the presence of a supplementary hidden layer of states that models data structure and captures the variability of the observations. HMMs have been successfully used in speech recognition (Jelinek 1976), image processing (Benmiloud and Pieczynski 1995), ecology (Le Ber et al. 2006) and landscape agronomy (Lazrak et al. 2010a).

An HMM2 is defined by three elements:

(i) A set $S = s_1, s_2,\ldots,s_N$ of N states. The states are the outcomes of the variables X_t, where $t = 1,\ldots,T$.

(ii) A transition matrix $A = (a_{ijk})$ over S^3, where a_{ijk} is the a priori transition probability $P(X_t = s_k/$

$X_{t-2} = s_i$, $X_{t-1} = s_j$) for the hidden Markov chain to be in state s_k at index t assuming it was in state s_j at index $t - 1$ and s_i at index $t - 2$. The Markov assumptions state that these a priori transition probabilities are constant.

(iii) A set of N distributions over a set of observations: $b_i(.)$ is the distribution of the observations associated with state s_i. The observations may be of different types: single land-use of a plot, several land-uses—called *n-uplet*—corresponding to the plot occupations during n successive years, or corresponding to the occupations of the n neighbouring plots. These distributions may be parametric: for example implemented in specific tables that store the observation probabilities, or represented by other HMMs that analyze an observation sequence at a whole and compute its probability. In this case, the HMM is called a hierarchical HMM. In our landscape clustering study based on land-use successions, we used a master HMM2 having six states, each of them being a 12 state HMM2. This latter analyzes the 12 year land-uses of a plot and computes the time-sequence probability. The master HMM has an ergodic topology: all the states are interconnected. The states describe the homogeneous areas (called patches) in the landscape. The transition probabilities account for the neighbourhood relations between patches. The master HMM models the spatial structure whereas the state HMM models the temporal structure. Such a model is capable of clustering a landscape into six patches whose evolution in terms of land-use successions is represented by a 12 state HMM. A more extended presentation of HMM2s, together with their performances in several data mining studies in agronomy and ecology can be found in Le Ber et al. (2006), Lazrak et al. (2010a).

Two separate and complementary data-mining analyses were conducted. In the first one, the observations are the land-uses of a plot. The regularities are revealed by a hierarchical HMM2 through the segmentation of the landscape into homogeneous patches, each of them having its land-use evolution described by a temporal state HMM2. In the second one, the observations are *n-uplets* elaborated from the land-uses of the neighbourhood plots. The regularities are revealed through the evolution of land-use neighbourhoods over time represented by a simple linear 6 state HMM2 that processes the *n-uplets*.

Data-mining software for identifying landscape regularities

ArpentAge[2] (Analyse de Régularités dans les Paysages: Environnement, Territoires, Agronomie = Landscape Regularities Analysis: Environment, Territories and Agronomy) is a software based on HMM2 for analyzing spatiotemporal data-bases (Lazrak et al. 2010a). ArpentAge takes as input an array of discrete data in which the columns contain the annual land-uses and the rows are regularly spaced locations of the studied landscape. The data-mining process starts with the data preparation, which consists of three stages: (i) defining land-use categories to reduce the great number of land-use modalities, (ii) defining the elementary observation (single category vs. *n-uplet* made of several land-use categories) and (iii) choosing the spatial resolution to sample the studied landscape (Lazrak et al. 2010a) because ArpentAge runs on raster data.

We simplified the 47 initial land-uses into 10 land-use categories as described by Lazrak et al. (2010a), with the slight difference that winter barley has been assigned to the "Wheat" category and the "Grassland and Alfalfa" category has been divided into two categories: "Grasslands" and "Alfalfa" (Table 1).

We sampled the landscape with regular spaced grids ranging from 10 m × 10 m up to 640 m × 640 m. Using each grid, we computed a feature: the number of different 12-year land-use successions. With a coarse resolution, small fields are omitted so that their land-use successions are lost. On the other hand, with a fine resolution, the huge matrix of sampled points does not allow tractable computations. We chose the method described in Lazrak et al. (2010a) to determine the grid resolution. A study of the variation of this feature as a function of the resolution showed that a 80 m grid resolution was a satisfying trade-off to avoid both long calculation times and the omission of small plot characteristics: only 6% of the 12-year land-use successions were lost.

[2] http://www.loria.fr/~jfmari/App/.

Table 1 Land-use categories used for data-mining analysis (time period: 1996–2007)

Land-use category	Land-use	Frequency	Cumul. frequency
Wheat (W)	Wheat, bearded wheat, winter barley, cereal[a]	0.372	0.372
Sunflower (S)	Sunflower, ryegrass followed by sunflower	0.139	0.511
Rapeseed (R)	Rapeseed	0.124	0.635
Urban (U)	Built area, peri-village, road	0.095	0.730
Maize (M)	Maize, rye grass followed by maize	0.076	0.806
Grasslands (G)	Permanent grassland, grassland first year, temporary grassland (2–3 years), grassland of unknown age	0.055	0.861
Forest and wasteland (F)	Forest or hedge, wasteland (uncultivated)	0.035	0.896
Alfalfa (A)	Alfalfa 1st year, alfalfa 2nd year, alfalfa 3rd year, alfalfa more than 3 years	0.026	0.922
Ryegrass (Y)	Ryegrass, ryegrass followed by ryegrass	0.024	0.946
Pea (P)	Pea	0.022	0.968
Others (O)	Spring barley, grape vine, spontaneous fallow in June, foxtail millet, flax, oat, clover, field bean, rye grass followed by tillage, rye grass followed by unknown, spontaneous fallow followed by tillage, rye, cereal-legume mixture, spring crop, mustard, garden/market gardening, sorghum/millet, sorghum, millet, tillage, tobacco, other crop	0.032	1.000

[a] Cereal is used when the species cannot be identified by the surveyor (it can be wheat, barley or other)

Land-use evolution data mining: clustering the Niort plain landscape into homogeneous patches

To perform the first data-mining analysis, we modelled the time-spatial structure of the landscape by a 6 state hierarchical HMM2 whose master HMM2 models the spatial structure, whereas the land-use evolution is modelled by the 12 state linear HMM2s (Fig. 1). The purpose of a linear HMM2 is to segment the study period in as many temporal segments as states (Mari and Le Ber 2006). In our case, each state was associated with one year. We located the land-use evolution by partitioning the study area into six homogeneous classes of patches in terms of land-use evolution. This value was obtained from previous studies on the same data (Lazrak et al. 2010a) and appeared to be a trade-off between heavy computations and useful clustering.

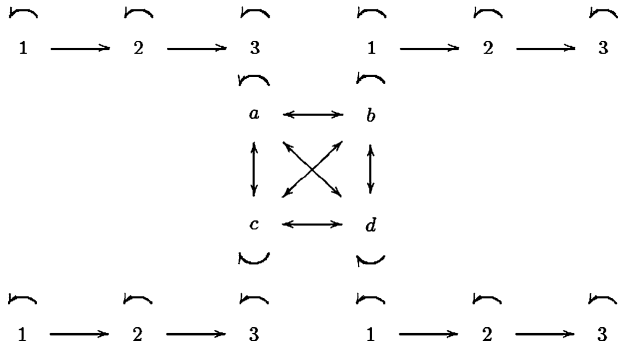


Fig. 1 Example of hierarchical HMM2. Each spatial state *a*, *b*, *c*, *d* of the master HMM2 (ergodic model) is a temporal HMM2 (linear model) the states of which are *1*, *2*, *3*. In our study, the ergodic model has six states, each of them being associated to a spatial area (patch), whereas the linear HMM2 have 12 states, each of them being associated to 1 year

Land-use neighbourhood data mining: analyzing the time evolution of neighbourhoods

In this second data-mining analysis, we aimed at representing more precisely the evolution of sunflower and maize neighbourhoods over time, since these crops are in jeopardy in the region due to their sensitivity to summer droughts. To do so, we explicitly considered their neighbourhood relationships in the studied landscape.

The elementary observation was a *5-uplet* of land-uses, also called quintuplet. A 80 m resolution gave too many neighbourhoods (approximately 40,000 different land-use quintuplets, more than the computer can process). We reduced this number by replacing land-use quintuplets by land-use cliques. A clique is a *2-uplet* made of the land-uses of two neighbouring

plots, regardless to their directions (Lazrak et al. 2010b). This replacement requires prior verification of the hypothesis that the land-use mosaic is isotropic (i.e. the direction does not hold any information). We studied the distribution of the *5-uplets* occurring in the 12-year study period and calculated the marginal joint probability $P(S, N)$ in each direction (North, South, East, West), where S and N are random variables of land-use categories, respectively in a site and in its neighbour site. For the studied landscape, we found that $P(S, N)$ were equal whatever the direction of neighbourhoods, which confirms that the land-use mosaic is isotropic and allows us to use the land-use cliques as elementary observations.

In order to assess the co-location of two land-use categories (x and y), we used Pointwise Mutual Information (PMI). PMI is usually used in text-mining (Novovičová et al. 2004) to find pairs of co-located words. It is defined as follows:

$$\mathrm{PMI}(x,y) = \log(P(\mathrm{N}=x, \mathrm{S}=y)/P(\mathrm{N}=x)*P(\mathrm{S}=y))$$

PMI compares the probability of observing x and y: together (joint probability) and independently (chance). If there is an attraction between x and y, the joint probability $P(x,y)$ is larger than $P(x)*P(y)$ and then PMI > 0 (Church and Hanks 1990). Inversely, if there is a repulsion between x and y, $P(x,y)$ is lower than $P(x)*P(y)$ and then PMI < 0. A zero value means that allocation to land of x is independent of the allocation of y since $P(x,y) = P(x)*P(y)$. PMI thus reveals attraction or repulsion regularities between land-use categories. Through on-farm surveys, agronomists may seek the decision rules explaining such attraction or repulsion regularities.

We here calculated the PMI on the basis of the cliques, regardless of their orientation. We then analysed the evolution of the neighbouring relationships of sunflower and maize by clustering the study period with a linear HMM2 of six states. This linear HMM2 defines six disjoint periods. In order to draw a global overview on neighbouring land-use evolution, we chose the periods 1998–2000 and 2004–2006 corresponding to states 2 and 5 of the 6-state linear HMM2.

By confronting the results of on-farm surveys about farmer decisions and the results of data-mining about landscape stochastic regularities, we now assess the coherence of the two approaches and how they can aid one another for landscape modelling.

Results

Farmer decision rules can be assessed at the landscape scale thanks to landscape stochastic regularities: example of sunflower

Through on-farm surveys from 2006 to 2010, we identified common farmer decision rules (between farmers) regarding sunflower. These rules referred to two decisional variables: (1) crop area and (2) suitable cultivation area. (1) After 2005, farmers generally chose to decrease the total area dedicated to sunflower on their farming territory (39 out of 67; Table 2). This decision was explained by two factors. The first one was the frequent summer droughts occurring in the region and particularly affecting sunflower yields, since this crop is planted in spring and needs water during summer. Farmers tended to replace sunflower by a less risky crop like rapeseed. The second factor was the European Common Agricultural Policy (CAP) reform of 2003 applied since 2006 in France. This factor specifically concerned mixed crop-livestock farms. Before the CAP reform, farmers got compensatory payments on the basis of the amount produced (product based subsidies): they had a specific amount of money for each cash crop, but no subsidy for grasslands. After the reform and the decoupling, they got single payments only depending on the eligible farm area (land based subsidies) and regardless of their cropping plan and production (Bougherara and Latruffe 2010). As a consequence, before the reform, most mixed crop-livestock farmers tended to grow rye-grass until May to cut it once for hay, but they planted sunflower just thereafter in order to get the annual subsidy. On the contrary, they now more frequently keep the planted rye-grass until fall to cut it several times in the year and get the subsidy even without sunflower.

(2) The decision of decreasing the sunflower area implied another common decision rule at farm scale (concerning the decisional variable "suitable cultivation area"): some farmers reduced the suitable cultivation area of sunflower on their farming territories and they concentrated it on the best places for sunflower. They stopped growing it close to the forests due to frequent damage of rabbits and crows in the vicinity of forests. In Table 2, we give the number of farmers having explicitly enunciated this rule (nine out of 26 asked farmers). One should however note that, unlike the nine farmers having enunciated the rule, 10 farmers out of 26 did not have several patches of forests within

Table 2 Farmer decision rules regarding sunflower and maize allocation to land (identified through on-farm surveys)

Decisional variable	Determinants	Decision rule	Number of farmers applying the rule				
			2006	2007	2009	2010	Total
Crop area for sunflower	Climatic risk affecting yields; CAP reform and decoupling	Reduce the sunflower area	12 (/22)	9 (/19)	7 (/12)	11 (/14)	39 (/67)
Suitable cultivation area for sunflower	Forests in the neighbourhood and associated crop damage	Plots close to forests (<500 m) not suitable for sunflower	–	–	3 (/12)	6 (/14)	9 (/26)
Suitable cultivation area for maize without irrigation	Type of soil (depth and humidity of soil)	Only plots with deep and humid soils suitable for maize without irrigation	9 (/22)	10 (/19)	9 (/12)	12 (/14)	40 (/67)
Crop area for maize on arable farms	Climatic risk affecting yield; only cash crop function (sold)	Reduce the maize area on arable farms	5 (/5)	8 (/10)	3 (/5)	3 (/6)	19 (/26)
Crop area for maize on mixed crop-livestock farms	Maize is used to feed herds	Maintain maize production on mixed crop-livestock farms	16 (/17)	8 (/9)	5 (/7)	6 (/8)	35 (/41)
Crop area for grasslands on mixed crop-livestock farms	Climatic risk affecting maize yields: need to secure fodder production	Increase the grasslands area on mixed crop-livestock farms	16 (/17)	9 (/9)	5 (/7)	5 (/8)	35 (/41)

a 500 m-distance of their plots or did not grow sunflower on their farms, and were thus not concerned by this rule.

We sought landscape stochastic regularities involving sunflower in order to see if the farmer decision rules identified at farm scale were consistent with observed landscape dynamics. First, the sunflower frequency substantially fell in the late 2000s at the landscape scale: the frequency was approximately divided by a factor of 2 between 1996 and 2007 (Fig. 2). Furthermore, Fig. 3a shows the evolution of the PMI between sunflower and other land-uses over the period and at landscape scale. It suggests that sunflower was in a relation of spatial repulsion with forests and grasslands over the whole period, while it was in a relation of spatial attraction with rapeseed and wheat. Besides, sunflower became less frequently close to forests and maize over the studied period. This landscape spatiotemporal regularity is thus consistent with individual farmer decision rules identified by on-farm surveys, which could explain the regularity at the landscape scale.

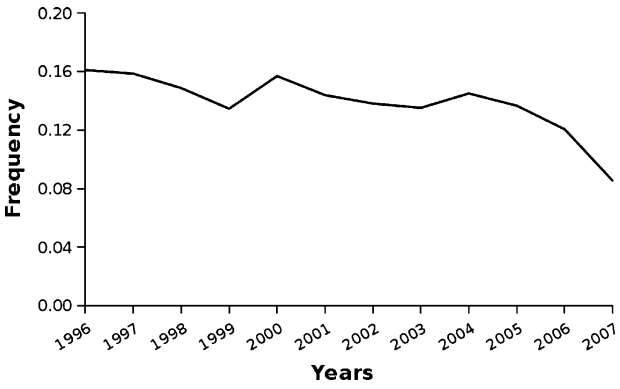


Fig. 2 Frequency evolution of sunflower at the landscape scale over the study period

Farmer decision rules contribute to explaining landscape stochastic regularities: example of maize

In parallel with the first example, Fig. 3b shows the evolution between the beginning (1998–2000) and the end (2004–2006) of the period of the PMI between maize and other land-uses at landscape scale. It clearly suggests that maize was more frequently close to grasslands (stronger attraction) and less frequently close to wheat (weaker attraction). It also shows that maize and respectively forests, sunflower and especially rapeseed were increasingly repulsed over the period.

We sought to identify generic farmer decision rules that would give explanatory elements to this time evolution of the neighbourhoods. Through on-farm surveys, we found that the maize area tended to decrease (decisional variable "crop area") and to be

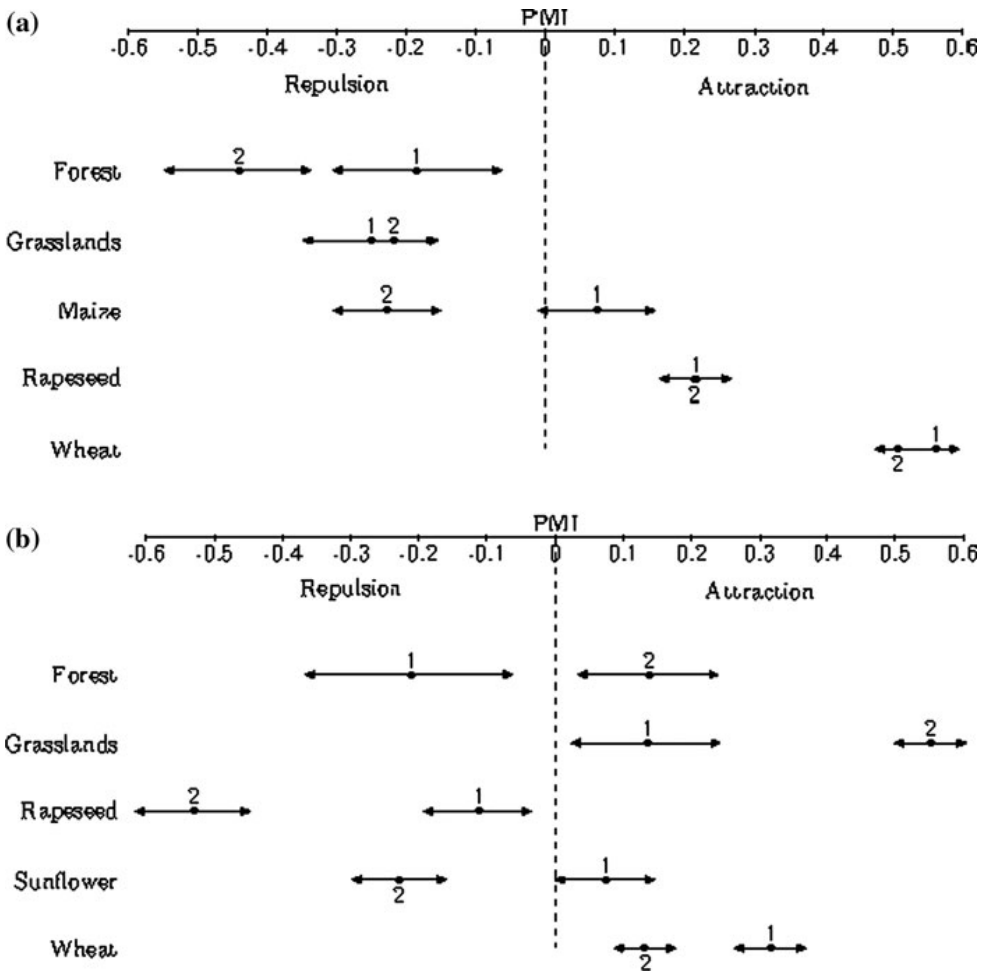


Fig. 3 PMI between sunflower and other land-uses (**a**) and between maize and other land-uses (**b**). For each land-use, two confidence intervals show the evolution of the neighbourhood starting in 1998–2000 (*1*) and ending in 2004–2006 (*2*). When the confidence intervals overlap themselves, only their union is shown. The interval shift shows the spatial attraction or repulsion process. The confidence intervals are computed using a 40 m resolution and a 5% risk. See Table 1 for land-use category details

mostly maintained in the vicinity of grasslands (decisional variable "suitable cultivation area") for several reasons. In fact, farmers had to adapt to a context of frequent summer droughts in the Niort plain landscape, leading to temporary irrigation bans (Martin et al. 2009). Maize (grain or silage) is only grown in deep humid soils, which are rather scarce in the Niort plain landscape (17% of arable soils) and/or in irrigated plots. Consequently, the adaptation strategies to water shortage mainly depend on the access to deep soils and the farming system. A common decision rule regarding the suitable cultivation area is that farmers only grow maize in deep and humid soils when they have no access to irrigation (40 out of 67; Table 2). They grow maize as a monoculture in the vicinity of grasslands traditionally well adapted to be located in deep and humid soils. 16 out of the 67 surveyed farmers were not concerned by the rule though, since they only grow corn in irrigated plots. Then, the crop allocation choices also depend on the farming system and the irrigation capacity, modulated by a risk of restriction. On arable farms, grain maize is not a priority crop and too risky due to irrigation bans, so that arable farmers decide to decrease the maize area (Table 2). On mixed crop-livestock farms, maize is a priority crop for feeding herds. Maize is thus mainly maintained as silage when they have access to irrigation. At the same time, maize production becomes increasingly risky due to climate variability and irrigation restrictions, so that mixed crop-livestock farmers tend to expand grasslands area to complement and secure fodder production (Table 2). This approach results in maize being mainly maintained on mixed crop-livestock farms with significant grasslands area on farms. Finally, all these decision rules appear to be generic and consistent with the fact that maize became more frequently close to grasslands over the period at landscape scale.

Links between time evolution of land-use neighbourhoods and clustering of the Niort plain landscape

After having commented the results concerning the analysis of the time evolution of land-use neighbourhoods, we now present the map resulting from the segmentation of the landscape using a Markovian framework (Fig. 4) and the possible consistency with previous results. This map consists of patches characterized by homogeneous evolution of land-use areas summarized in small associated graphs. The map thus gives the possibility to locate, inside the landscape, the places where a certain land-use area decreased or increased and to compare the patches regarding their land-use area evolutions over time. The map unit (a) refers to patches where maize is the most frequent land-use. In these patches, grasslands seem to have increased since 2004 simultaneously to maize decrease. As a consequence, these patches may represent mixed crop-livestock farmer practices, as described above. The map units (b) and (c) represent crop areas where maize has been partially replaced by grasslands. The map unit (d) refers to patches where grasslands are the most frequent land-use, which may correspond to areas with deep and humid soils. Despite a global decreasing trend, maize appears to be more frequent in this grassland patch than in crop areas, which is consistent with the rule consisting in growing maize only in deep and humid soils. Map unit (e) represents patches were forests are the most important land-use and where sunflower is very scarce and decreasing over the period.

Discussion

A new framework for landscape modelling

We have here pointed out that the two modelling methods (farmer decision rules analysis and landscape stochastic regularities computation) aid one another for land-use modelling at the landscape scale and understanding the driving forces of its spatial organization. In line with Thenail et al. (2009) and Sorel et al. (2010), we argue that spatiotemporal crop allocation to field patterns is designed at the farm scale and that mainly farmers design agricultural landscapes, which makes the analysis of their decisions more crucial compared to other stakeholders.

We managed to create a map of spatiotemporal regularities (Fig. 4), partly explained by generic farmer decision rules, mostly referring to the decisional variables "crop area" and "suitable crop areas". The combination of an ergodic HMM and a Markov chain made it possible to account for both spatial and temporal changes, despite the difficulties of handling both these dimensions (Verburg 2006). Following the work of Lazrak et al. (2010a), this map constitutes a partitioning of the landscape, based on both its spatial and temporal organization and maximizing the probability that the model fits the data. This drawing of new spatial patterns coherent with land-use dynamics at the landscape scale may improve the links to the scale of ecological processes (Pelosi et al. 2010). We could in addition account for possible changes in crop proportions whereas simple transition matrices in Markov chains (Castellazzi et al. 2008) induce stationary crop proportions when used to simulate crop successions.

Moreover, in contrast to a time-invariance of the socioeconomic context and of the driving factors of landscape changes (Sorel et al. 2010), an originality of our approach relies on accounting for a changing context. The modification of socioeconomic and climatic driving factors induced in our case changes in crop proportions over the studied period. In the Niort plain region, the EU CAP reform contributed to decreasing the total sunflower area, while the frequent drought risk contributed to reducing the maize area. These crop proportion changes resulted in changes in the landscape patterns, which can in turn impact environmental issues. Hence, our approach seems to be useful for landscape modelling and thus for a better knowledge of the interactions between ecological processes and landscape dynamics.

Remaining challenges

One limit of our approach is that not all farmer decisions are generic enough to be assessed at landscape scale. For example, through on-farm surveys, we identified two different management strategies for alfalfa. On arable farms, a common rule was to implant alfalfa all at the same time in order to simplify the cropping system and only in marginal areas. It was also a mean for farmers to get specific subsidies within the framework of the CAP Territorial Agroenvironmental Measures implemented in France. On the

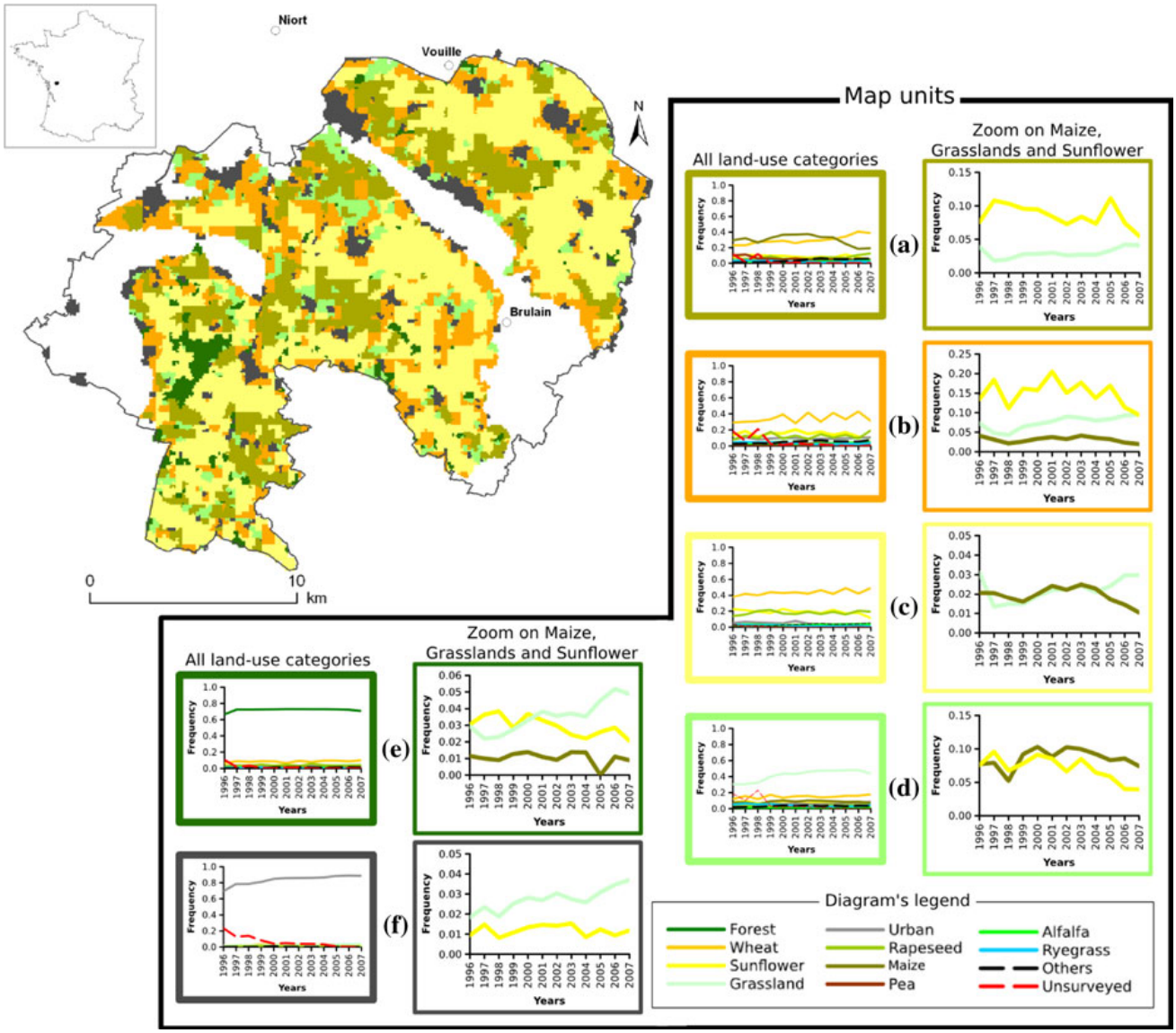


Fig. 4 Segmentation of the Niort Plain landscape in patches characterized by homogeneous evolution of land-use areas over the studied period. *White areas* are unclassified because there was insufficiently surveyed land-use over the 1996–2007 period. The location of the Niort Plain in France is depicted in the *upper left-hand box*. The map legend (evolution of land-use areas in each patch) is illustrated in small *graphs* in the frame. There are six map units: (**a**)–(**f**). Each *map* unit is described by two diagrams: (i) a *left-hand diagram* showing the evolution of all land-use categories in patches belonging to the considered map unit and (ii) a *right-hand diagram* showing a zoom on interesting land-use categories. The map unit (**a**) refers to patches where maize is the most frequent land-use. The map units **b**, **c** represent crop areas where maize has been partially replaced by grasslands. The map unit **d** refers to patches where grasslands are the most frequent land-use. Map unit **e** represents patches were forests are the most important land-use. Map unit **f** represents urban areas

contrary, on mixed crop-livestock farms, a common rule was to grow alfalfa of different ages (1–5 years). The gradual implantation aims at securing the fodder production considering that alfalfa yields depend on the age of the implanted alfalfa. And yet, it was not possible to identify a spatiotemporal regularity about alfalfa. The fact that alfalfa concerns a marginal area (<4%) compared to commercial crops can explain the difficulty in identifying landscape regularity and assessing the associated rule at landscape scale. Sorel et al. (2010) also noted that marginal crops (the ones with small proportions of the landscape area) were the least well predicted concerning crop spatiotemporal allocation compared to major crops. This observation confirms that stochastic modelling, either Markov models or stochastic decision trees, is difficult to carry

out for marginal areas. Nevertheless, even if not all generic, we think that it is still important to identify these decisions because they give the possibility of identifying innovative and potential future farmer adaptations to a changing context.

Moreover, as mentioned by Thenail and Baudry (2004), many decisions for spatiotemporal crop allocation to landscape patterns are specific to certain farm types (e.g. arable vs. mixed crop-livestock farms). Given that the different farm types can be spread over discontinuous landscape units within the landscape, it is difficult to assess specific rules at the landscape scale. This is again consistent with the results of Sorel et al. (2010): they appear to get less success in predicting spatiotemporal crop allocation when using generic Markov models than when using farm type specific ones. In our approach, we did not identify stochastic regularities at the level of a discontinuous landscape composed by only a certain farm type: our interest was to link the farm and the landscape scales when farmer rules are generic enough to impact landscape patterns. One must however notice that the choice of considering continuous or discontinuous landscapes is of variable interest depending on the ecological process studied and the objective to be achieved (Pelosi et al. 2010).

A second limit of our approach is due to the difficulty in determining the part of each rule explaining a regularity when several farmer decision rules are possible. For example, we identified a strong regularity of neighbourhood between grasslands and built-up areas (not shown). Two possible farmer rules could explain this regularity. The first one is that most farmers allocate permanent grass or set-asides instead of commercial crops next to houses in order to avoid agrochemical spraying in small plots close to citizens' houses. The second rule is that most dairy breeders put grasslands just next to the milking room so that dairy cows can graze in the vicinity of the dairy barn, which is in accordance with several authors (Benoît 1990; Marie et al. 2009). Our difficulty is to determine the part of each rule in explaining the regularity of the neighbourhood between grasslands and built-up areas, which is also related to the scarcity of available data: there is indeed no data-base about buildings in rural landscapes with a distinction between urban and agricultural buildings (and among them, with the specific use of each building). To overcome the difficulty of determining the explaining part of different rules, one interesting perspective of our work could be the complementary use of landscape simulation software. We could generate landscape patterns according to different farmer rules and compare them to random landscapes vs. real landscapes.

Conclusion: contribution of landscape agronomy to landscape ecology

As a conclusion, the two modelling methods of farmer decisions and landscape regularities have been respectively used for a long time now. Our paper suggests that a new approach consisting in a combination of the two methods helps in articulating the farm and the landscape scales for land-use modelling and improving our understanding of land-use processes. The originalities rely on (i) the combination of two methods used separately so far and (ii) the accounting for both the spatial and the temporal dimensions of crop allocation to landscape patterns in a changing context. Thanks to more on-farm surveys and to remote sensing improvements and developing spatial land-use data-bases (e.g. CAP declarations in Europe) for data-mining, such an approach could be applied in the future in other landscapes for upscaling.

As landscape agronomists, we consider that the crop mosaic inside agricultural landscapes is organized by farmers. On the other hand, landscape ecologists view the landscape as a random-like phenomenon influenced by natural factors (Burel and Baudry 2010). Our study therefore seeks to contribute to landscape ecology through a deeper insight into the relationships between LSO, its driving forces and its impacts on ecological processes.

Acknowledgments We thank the CEBC for the land-use data-base. We are also very grateful to Benoît Lelaure, Gaëdig Méola and Camille Bernard, who carried out the farmer surveys in 2006, 2007 and 2010. The PhD of N. Schaller was funded by the Ile-de-France region through the DIM ASTREA and the PhD of E.G. Lazrak was funded by the Lorraine Region and the ANR BioDivAgriM project. We would like to sincerely thank Donald White for improving the English, Aude Barbottin and the three anonymous reviewers for their fruitful comments.

References

Aubry C, Biarnes A, Maxime F, Papy F (1998a) Modélisation de l'organisation technique de la production dans

l'exploitation agricole : la constitution de système de culture. Etud Rech Syst Agraires Dév 31:25–43
Aubry C, Papy F, Capillon A (1998b) Modelling decision-making processes for annual crop management. Agric Syst 56(1):45–65
Aubry C, Paillat J-M, Guerrin F (2006) A conceptual model of animal wastes management in the Reunion Island. Agric Syst 88:294–315
Bacic I, Rossiter D, Bregt A (2006) Using spatial information to improve collective understanding of shared environmental problems at watershed level. Landsc Urban Plan 77:54–66
Beaujouan V, Durand P, Ruiz L (2001) Modelling the effect of the spatial distribution of agricultural practices on nitrogen fluxes in rural catchments. Ecol Model 137:93–105
Benmiloud B, Pieczynski W (1995) Estimation des paramètres dans les chaînes de Markov cachés et segmentation d'images. Trait Signal 12(5):433–454
Benoît M (1990) La gestion territoriale de l'activité agricole dans un village lorrain. Mappemonde 4:15–17
Benoît M, Deffontaines JP, Gras F, Bienaimé E, Riela-Cosserat R (1997) Agriculture et qualité de l'eau. Une approche interdisciplinaire de la pollution par les nitrates d'un bassin d'alimentation. Cah Agric 6:97–105
Benoît M, Mignolet C, Hermann S, Rizzo D, Moonen C, Barberi P, Galli M, Bonari E, Silvestri N, Thenail C, Lardon S, Rapey H, Marraccini E, Le Ber F, Meynard JM (2007) Landscape as designed by farming systems: a challenge for landscape agronomists in Europe. In: Farming systems design 2007, methodologies for integrated analysis of farm production systems, Catania, pp 137–138
Benton TG, Vickery JA, Wilson JD (2003) Farmland biodiversity: is habitat heterogeneity the key? Trends Ecol Evol 18(4):182–188
Bougherara D, Latruffe L (2010) Potential impact of the EU 2003 CAP reform on land idling decisions of French landowners: results from a survey of intentions. Land Use Policy 27(4):1153–1159
Brunschwig G, Josien E, Bernhard C (2006) Contraintes géographiques et modes d'utilisation des parcelles en élevage bovin laitier et allaitant. Fourrages 185:83–95
Burel F, Baudry J (2010). Landscape and resilience. In: Proceedings of 'Agro2010 the XIth ESA congress', Montpellier, France, pp 143–144
Castellazzi MS, Perry JN, Colbach N, Monod H, Adamczyk K, Viaud V, Conrad KF (2007) New measures and tests of temporal and spatial pattern of crops in agricultural landscapes. Agric Ecosyst Environ 118(1–4):339–349
Castellazzi MS, Wood GA, Burgess PJ, Morris J, Conrad KF, Perry JN (2008) A systematic representation of crop rotations. Agric Syst 97(1–2):26–33
Church KW, Hanks P (1990) Word association norms, mutual information, and lexicography. Comput Linguist 16(1):22–29
Cumming G, Cumming DHM, Redman CL (2006) Scale mismatches in social-ecological systems: causes, consequences, and solutions. Ecol Soc 11(1):14
de Koning GHJ, Verburg PH, Veldkamp A, Fresco LO (1999) Multi-scale modelling of land use change dynamics in Ecuador. Agric Syst 61:77–93
Donald P, Green RE, Heath MF (2001) Agricultural intensification and the collapse of Europe's farmland bird populations. Proc R Soc Lond 268:25–29
Dury J, Schaller N, Garcia F, Reynaud A, Bergez JE (2011) Models to support cropping plan and crop rotation decisions. A review. Agron Sustain Dev (Online First)
Freeman T, Nolan J, Schoney R (2009) An agent-based simulation model of structural change in Canadian Prairie agriculture, 1960–2000. Can J Agric Econ 57:537–554
Gaucherel C, Houet T (2009) Preface to the selected papers on spatially explicit landscape modelling: current practices and challenges. Ecol Model 220:3477–3480
Happe K, Schnicke H, Sahrbacher C, Kellermann K (2009) Will they stay or will they go? Simulating the dynamics of single-holder farms in a dualistic farm structure in Slovakia. Can J Agric Econ 57:497–511
Havet A, Martin P, Laurent M, Lelaure B (2010) Adaptation des exploitations laitières aux incertitudes climatiques et aux nouvelles réglementations. Le cas des productions bovines et caprines en Plaine de Niort. Fourrages 202: 145–151
Jelinek F (1976) Continuous speech recognition by statistical methods. Proc IEEE 64:532–556
Joannon A, Souchère V, Martin P, Papy F (2006) Reducing runoff by managing crop location at the catchment level, considering agronomic constraints at farm level. Land Degrad Dev 17(5):467–478
Joannon A, Bro E, Thenail C, Baudry J (2008) Crop patterns and habitat preferences of the grey partridge farmland bird. Agron Sustain Dev 28:379–387
Lambin EF, Geist HJ, Lepers E (2003) Dynamics of land-use and land-cover change in tropical regions. Annu Rev Environ Resour 28:205–241
Landais E (1998) Modelling farm diversity. New approaches to typology building in France. Agric Syst 58(4):505–527
Lazrak EG, Mari JF, Benoît M (2010a) Landscape regularity modelling for environmental challenges in agriculture. Landscape Ecol 25:169–183
Lazrak EG, Benoît M, Mari JF (2010b) Time-space dependencies in land-use successions at the scale of an agricultural landscape. In: International conference on integrative landscape modelling—LandMod 2010. Symposcience
Le Bail M, Lecroart B, Gauffreteau A, Angevin F, Messean A (2010) Effect of the structural variables of landscapes on the risks of spatial dissemination between GM and non-GM maize. Eur J Agric 33:12–23
Le Ber F, Benoît M (1998) Modelling the spatial organization of land use in a farming territory. Example of a village in the Plateau Lorrain. Agronomie 18(2):103–115
Le Ber F, Benoît M, Schott C, Mari JF, Mignolet C (2006) Studying crop sequences with CarrotAge, a HMM-based data mining software. Ecol Model 191(1):170–185
Leisz SJ, Thu Ha NT, Bich Yen NT, Thanh Lam N, Duc Vien T (2005) Developing a methodology for identifying, mapping and potentially monitoring the distribution of general farming system types in Vietnam's northern mountain region. Agric Syst 85:340–363
Mari J-F, Le Ber F (2006) Temporal and spatial data mining with second-order Hidden Markov models. Soft Comput 10:406–414
Marie M, Bensaid A, Delahaye D (2009) Le rôle de la distance dans l'organisation des pratiques et des paysages agricoles : l'exemple du fonctionnement des exploitations laitières dans l'arc atlantique. Cybergeo: Eur J of Geography.

Cartographie, Imagerie, SIG, article 460. URL:http://cybergeo.revues.org/index22366.html

Martin P, Schaller N, Havet A (2009) Diversity of farmers' adaptations to a new context of irrigation restrictions: consequences on grassland area development. In: Proceedings of the symposium farming system design, Monterey, CA, pp 249–250

Mawois M, Aubry A, Le Bail M (2011) Can farmers extend their cultivation areas in urban agriculture? A contribution from agronomic analysis of market gardening systems around Mahajanga (Madagascar). Land Use Policy 28(2):434–445

Maxime F, Mollet JM, Papy F (1995) Aide au raisonnement de l'assolement en grande culture. Cah Agric 4:351–362

Merot A, Bergez JE, Capillon A, Wery J (2008) Analysing farming practices to develop a numerical, operational model of farmers' decision-making processes: an irrigated hay cropping system in France. Agric Syst 98(2):108–118

Mignolet C, Schott C, Benoît M (2007) Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale. Sci Total Environ 375:13–32

Morlon P, Benoît M (1990) Étude méthodologique d'un parcellaire d'exploitation agricole en tant que système. Agronomie 6:499–508

Morlon P, Trouche G (2005) Nouveaux enjeux de la logistique dans les exploitations de grande culture. L'organisation spatiale des assolements : exemple et questions. Cah Agric 14(3):305–311

Navarrete M, Le Bail M (2007) SALADPLAN: a model of the decision-making process in lettuce and endive cropping. Agron Sust Dev 27(3):209–221

Novovičová J, Malik A, Pudil P (2004) Feature selection using improved mutual information for text classification. In: Fred A, Caelli T, Duin RPW, Campilho A, Ridder D (eds) Structural, syntactic, and statistical pattern recognition, vol 3138 of Lecture notes in computer science. Springer, Berlin, pp 1010–1017

Pelosi C, Goulard M, Balent G (2010) The spatial scale mismatch between ecological processes and agricultural management: do difficulties come from underlying theoretical frameworks? Agric Ecosyst Environ 139:455–462

Pocewicz A, Nielsen-Pincus M, Goldberg CS, Johnson MH, Morgan P, Force JE, Waits LP, Vierling L (2008) Predicting land use change: comparison of models based on landowner surveys and historical land cover trends. Landscape Ecol 23:195–210

Rindfuss RR, Walsch SJ, Turner BL II, Fox J, Mishra V (2004) Developing a science of land change: challenges and methodological issues. PNAS 101(39):13976–13981

Robinson DT, Brown DG (2009) Evaluating the effects of land-use development policies on ex-urban forest cover: an integrated agent-based GIS approach. Int J Geogr Inf Sci 23(9):1211–1232

Rounsevell MDA, Annetts JE, Audsley E, Mayr T, Reginster I (2003) Modelling the spatial distribution of agricultural land use at the regional scale. Agric Ecosyst Environ 95(2–3):465–479

Schaller N, Aubry C, Martin P (2010) Modelling farmers' decisions of splitting agricultural plots at different time scales: a contribution for modelling landscape spatial configuration. In: Proceedings of 'Agro2010 the XIth ESA congress', Montpellier, France, pp 879–880

Sebillotte M, Soler LG (1990) Les processus de décision des agriculteurs : acquis et questions vives. In: Brossier J, Vissac B, Lemoigne JL (eds) Modélisation systémique et systèmes agraires. INRA, Paris, pp 93–102

Sorel L, Viaud V, Durand P, Walter C (2010) Modeling spatio-temporal crop allocation patterns by a stochastic decision tree method, considering agronomic driving factors. Agric Syst 103:647–655

Stoate C, Boatman ND, Borralho RJ, Rio Carvalho C, de Snoo GR, Eden P (2001) Ecological impacts of arable intensification in Europe. J Environ Manag 63(4):337–365

Stoate C, Baldi A, Beja P, Boatman ND, Herzon I, van Doorn A, de Snoo GR, Rakosy L, Ramwell C (2009) Ecological impacts of early 21st century agricultural change in Europe—a review. J Environ Manag 91:22–46

Thenail C, Baudry J (2004) Variation of farm spatial land use pattern according to the structure of hedgerow network (bocage) landscape: a case study in Northern Brittany. Agric Ecosyst Environ 101:53–72

Thenail C, Joannon A, Capitaine M, Souchère V, Mignolet C, Schermann N, Di Pietro F, Pons Y, Gaucherel C, Viaud V, Baudry J (2009) The contribution of crop-rotation organization in farms to crop-mosaic patterning at local landscape scales. Agric Ecosyst Environ 131:207–219

Turner M (1990) Spatial and temporal analysis of landscape pattern. Landscape Ecol 4:21–30

Valbuena D, Verburg PH, Bregt AK, Ligtenberg A (2010) An agent-based approach to model land-use change at a regional scale. Landscape Ecol 25:185–199

van Oost K, Govers G, Desmet P (2000) Evaluating the effects of changes in landscape structure on soil erosion by water and tillage. Landscape Ecol 15:577–589

Veldkamp A, Fresco LO (1996) CLUE-CR: an integrated multi-scale model to simulate land use change scenarios in Costa Rica. Ecol Model 91:231–248

Veldkamp A, Lambin EG (2001) Predicting land-use change: editorial. Agric Ecosyst Environ 85:1–6

Verburg PH (2006) Simulating feedbacks in land use and land cover change models. Landscape Ecol 21:1171–1183

Chapitre 7

Discussion Générale

7.1 Retour sur les hypothèses

7.1.1 L'espace agricole représenté par sa mosaïque de successions d'occupation du sol

Nous avons modélisé l'Organisation Territoriale de l'Activité Agricole (OTAA) en étudiant l'organisation spatiale des systèmes de culture et leurs dynamiques. Nous avons décrit ces systèmes de cultures uniquement par leur première composante : la succession de cultures (SEBILLOTTE, 1990) sans introduire les itinéraires techniques. Par contre, nous avons élargi notre étude aux surfaces non agricoles, en élargissant le concept de succession de culture à celui de la succession d'occupation du sol (OCS). Nous avons assimilé l'OTAA à un champ markovien de successions d'OCS où l'évolution du processus à un instant ou (et) à un lieu donné(s) est uniquement déterminée par les valeurs du voisinage temporel ou (et) spatial. Cette représentation suppose que :

1. l'OCS d'une année dépend des cultures au même endroit de quelques années précédentes ;
2. les OCS sont organisées dans l'espace et dépendent de leur voisinage proche ;
3. le choix d'une succession d'OCS en un point de l'espace dépend du choix des successions des cultures voisines.

Notre représentation de l'OTAA en champ markovien de successions d'OCS est une implémentation agronomique du paysage agricole vu par les écologues du paysage comme une mosaïque de types d'usages du sol (BONTHOUX, 2011). La succession d'OCS étant l'élément clé que nous avons exploité pour expliciter le type d'usage du sol dans les espaces agricoles.

Dès la fin du XIX^{e} siècle, le concept de succession végétale a été introduit pour l'étude de la dynamique de la végétation dans des écosystèmes naturels (GLENN-LEWIN et VAN DER MAAREL, 1992), et les modèles de Markov ont rapidement prouvé leur utilité pour ce genre d'études (USHER, 1992). Plus

récemment, LADET et al. (2005) ont développé une application SIG fondée sur des chaînes de Markov spatialisées avec un automate cellulaire permettant de prédire les évolutions futures des dynamiques paysagères à partir de la connaissance des dynamiques antérieures. Dans un objectif similaire de simulation de paysages, CASTELLAZZI et al. (2008) ont développé une application [1], qui s'appuie sur une modélisation markovienne pour simuler l'organisation des rotations de cultures à l'échelle régionale. Ces travaux cités à titre indicatif et non exhaustif témoignent de l'intérêt de la modélisation markovienne dans l'étude des dynamiques végétales qu'elles soient naturelles ou agricoles. Leur point commun est qu'ils implémentent des chaînes de Markov dont la subjectivité liée à l'estimation par expertise de leurs paramètres en limite considérablement l'intérêt. Les HMM généralisent les chaînes de Markov et présentent l'avantage de posséder des algorithmes d'estimation automatique de leurs paramètres à partir d'un corpus de données (BAUM et al., 1970 ; WELCH, 2003). Les HMM ont été initialement développés pour reconnaître la parole (JELINEK, 1976), ils ont ensuite été adaptés pour reconnaitre les formes dans des images numériques (BENMILOUD et PIECZYNSKI, 1995). À notre connaissance, l'utilisation des HMM pour modéliser les dynamiques d'usages du sol dans des territoires agricoles se limite aux travaux de notre équipe de recherche. L'hypothèse « anthropomorphique » de nos travaux a été de comparer une région à une personne qui parle et de postuler que les HMM allaient nous permettre de comprendre ce que les successions d'OCS faisaient dire à une région agricole tout comme en reconnaissance de la parole où les HMM sont utilisés pour identifier des mots ou des phonèmes dans la suite de sons élémentaires produits par un conduit vocal.

Enfin, cette thèse se démarque des travaux antérieurs de l'équipe par une prise en compte explicite et couplée des composantes temporelle et spatiale de l'OTAA. Notre représentation de l'OTAA en champ de Markov de successions d'OCS attribue le même poids à ces deux composantes, et constitue un cadre conceptuel pertinent pour utiliser un seul outil — les HMM — pour modéliser les aspects temporels et spatiaux de l'OTAA.

7.1.2 L'approche temporo-spatiale est efficace pour modéliser l'OTAA

Notre représentation de l'OTAA comme un champ de Markov de successions d'OCS assimile l'organisation temporelle et spatiale des OCS à un couplage de chaînes de Markov et de champs de Markov qui définit les dépendances temporelles dans les successions d'OCS et les dépendances spatiales de localisation de ces successions. Une conséquence de cette représentation et un résultat important de notre thèse est la mise au point d'une approche de modélisation de l'OTAA qui traite dans l'ordre le temps avant l'espace pour ensuite localiser les régularités temporelles identifiées.

Cet ordre dans notre démarche trouve sa raison dans la nature des dynamiques des territoires agricoles où s'articulent « temps rond » et « temps long » (DEFFONTAINES et al., 1997) :

– un « temps rond » d'une activité agricole qui suit une cyclicité dictée par les assolements et les rotations,

1. LandSFACTS : *cf.* http ://www.macaulay.ac.uk/LandSFACTS/model.php.

– et un « temps long » pendant lequel s'opère le changement proprement dit. Ce changement se traduit par des successions faisant la transition vers des systèmes de cultures stables.

Qu'il s'agisse de « temps rond » ou de « temps long », l'activité agricole génère des successions de cultures qui renferment des logiques agronomiques (SEBILLOTTE, 1974) qu'il convient d'analyser et de comprendre avant de s'occuper de leurs organisation spatiale. Cette particularité des dynamiques des territoires agricoles est le fondement de notre choix méthodologique consistant à traiter le temps avant l'espace, c'est aussi la raison de son efficacité, car dans un territoire agricole le changement dans le temps et (ou) dans l'espace n'est pour majorité qu'apparent. L'idée à terme est de pouvoir distinguer :

– les rotations qui se pratiquent pendant les périodes de stabilité des systèmes de cultures (« temps rond »), où les successions prennent alors ces formes régulières et particulières,
– des successions d'OCS qui marquent des transitions vers de nouveaux systèmes de cultures et qui traduisent l'adaptation des agriculteurs aux contraintes et opportunités qu'ils rencontrent.

Devant la difficulté à automatiser cette distinction, nous avons choisi de nous intéresser aux successions d'OCS (LAZRAK et al., 2009).

Enfin, cette approche de modélisation que nous avons qualifiée de temporo-spatiale, permet de cartographier des objets temporels préalablement identifiés. Elle est différente de la démarche spatio-temporelle classique qui suit les trajectoires d'objets spatiaux au travers d'une série temporelle de cartes ou plus généralement d'images (SIMPSON et al., 1994 ; POUDEVIGNE et ALARD, 1997). Cette dernière approche est peut-être adaptée à l'étude de la détection de mouvements (CAPLIER, 1995) ou à l'étude des dynamiques urbaines (SIMON, 2012), mais n'est pas adaptée à l'étude des dynamiques des territoires agricoles (PAPY et TORRE, 2002 ; CREWS-MEYER, 2004 ; MARTIN et al., 2006 ; BENOÎT, 2006 ; MIGNOLET et al., 2007 ; SCHIETTECATTE et al., 2008 ; THENAIL et al., 2009). Nous défendons l'idée que la modélisation de l'OTAA par la cartographie des dynamiques temporelles — telles que des successions d'OCS (LAZRAK et al., 2009), ou des évolutions d'assolements (SCHALLER et al., 2011) — est plus pertinente que l'approche classique consistant à représenter les dynamiques agricoles par l'analyse d'une succession d'images.

7.1.3 Notre méthode de modélisation est applicable aux données de télédétection

Dans le cadre de cette thèse, nous avons montré que notre méthode de modélisation, initialement développée pour des données vectorielles issues de relevés de terrain, peut traiter des bases de données spatio-temporelles d'OCS issues de télédétection, si celles-ci sont en mesure d'informer une longue période avec un pas de temps annuel. La figure 7.1 illustre l'application de cette méthode de modélisation sur les données de télédétection à travers l'exemple du bassin versant du Yar (Bretagne, France) où l'activité agricole est à l'origine de problèmes d'eutrophisation dans la baie de Saint-Michel-en-Grève. Nous avons réalisé une segmentation temporo-spatiale sur le corpus de données d'OCS du bassin versant du Yar. Cette segmentation a permis de partitionner le bassin versant du Yar en zones homogènes d'évolution des OCS. Ce partitionnement décrit l'OTAA dans ce bassin versant de manière simple et révèle des dynamiques

cachées pouvant contribuer à mieux comprendre l'implication de l'OTAA dans l'eutrophisation de la baie de Saint-Michel-en-Grève.

Notre méthode de modélisation exige en entrée un corpus de données qui jusqu'à présent a été fourni à partir de deux sources : les relevés annuels de terrain, et la télédétection.

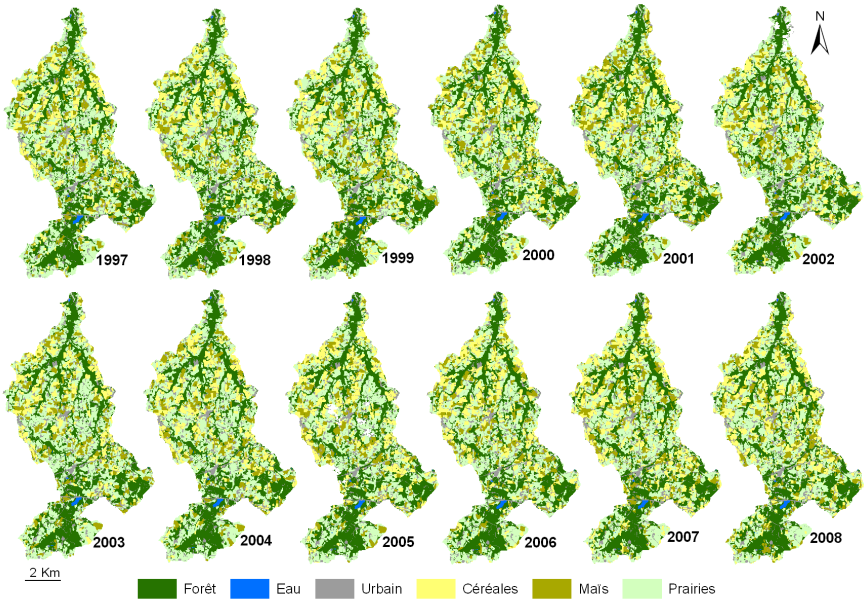
La construction à l'échelle d'un territoire régional et le maintien sur du long terme d'une base de données issue de relevés réguliers de terrain comme celle du CEBC (*cf.* section 2.1.2 page 19) est sans doute anecdotique vue les énormes moyens humains que ceci nécessite. Elle est aussi susceptible de se dégrader (par exemple Enquête *Teruti-Lucas* non réalisée en 2011).

La télédétection ouvre une alternative prometteuse que nous avons explorée à travers la modélisation de l'OTAA dans le bassin versant du Yar. La télédétection est devenue une source permettant de construire des bases de données spatio-temporelles d'OCS dont la qualité a été grandement améliorée au cours des dernières années. Les bases de données d'OCS issues de traitements d'images de télédétection présentent, toutefois, des inconvénients par rapport à celles issues de relevés de terrain. En particulier, les images de télédétection offrent une diversité thématique plus faible (peu de modalités d'OCS), elles n'identifient pas les limites parcellaires et sont sujettes à une faible sécurité de saisie en raison, en l'occurrence, du risque d'ennuagement au moment du passage du satellite sur le territoire d'étude. Enfin, les progrès envisagés par les nouveaux satellites (comme Sentinelle) permettent d'espérer la production de cartes d'OCS thématiquement plus riches à des échelles de territoires régionaux avec la reconnaissance des limites des parcelles et des mises à jour fréquentes des informations, mais il faudra attendre une dizaine d'années avant de pouvoir disposer d'une base de données avec une telle qualité.

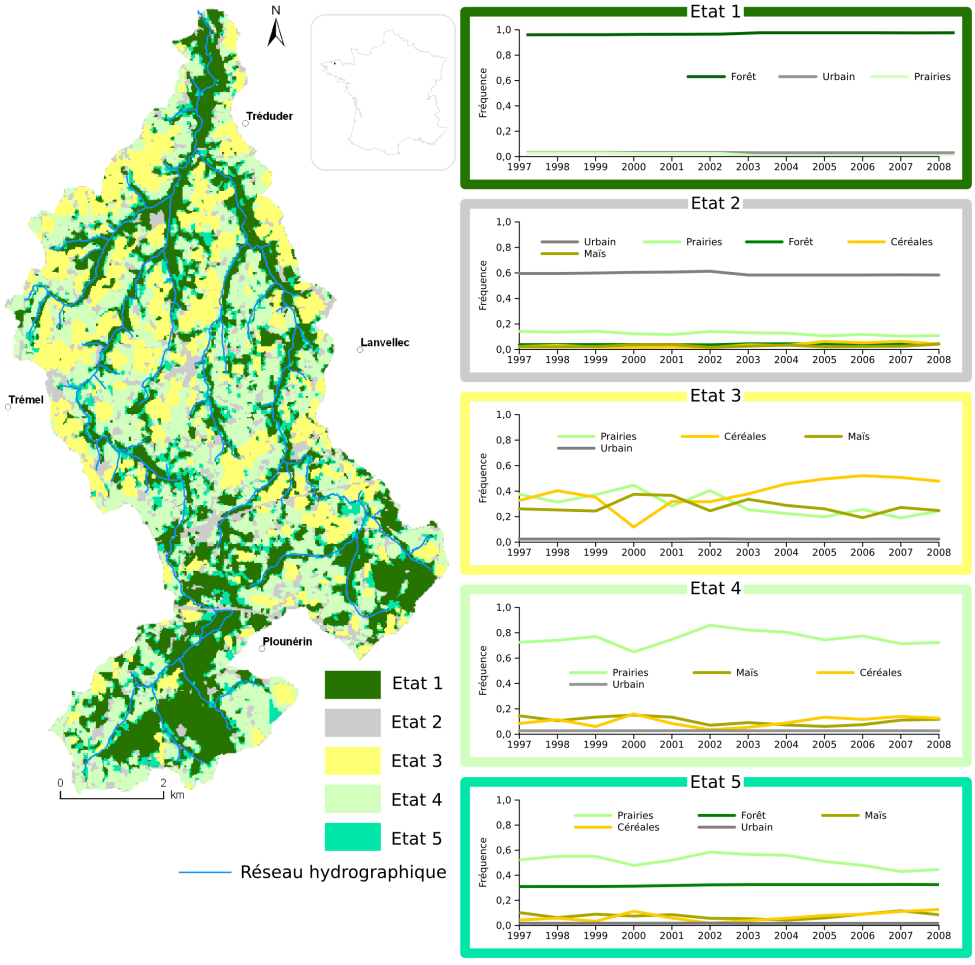
En conclusion, le corpus de données mobilisable reste le facteur limitant de la mise en œuvre de notre méthode de modélisation. Une autre alternative qui permettrait de construire des bases de données spatio-temporelles d'OCS avec les qualités recherchées consisterait à réaliser des enquêtes de terrain exhaustives auprès des exploitations agricoles (EA) du territoire d'étude, avec un rythme pluriannuel, en consultant les archives des EA et la mémoire des agriculteurs (MOTTET et al., 2006). Cette dernière alternative, bien que plus lourde à mener, est au moins aussi prometteuse que la télédétection car elle permettrait d'obtenir des informations plus précises sur les cultures, sur les itinéraires techniques et sur les règles de décisions d'assolements opérés par les agriculteurs.

7.1.4 L'OTAA est organisée de manière convergente et intelligible

Partant du constat que l'OTAA résulte en grande partie de l'activité des agriculteurs qui façonnent le territoire agricole chacun au niveau de son exploitation (DEFFONTAINES et al., 1995 ; DEFFONTAINES et THINON, 2001 ; THENAIL et al., 2009), nous avions fait l'hypothèse que l'activité des différents agriculteurs, sans être concertée, conduit à l'organisation de l'espace agricole d'une manière convergente et intelligible que nous avons exploitée dans notre modélisation de l'OTAA. L'hypothèse de convergence nous a permis de chercher à représenter l'OTAA en zones compactes (patches) caractérisées par des dynamiques agricoles semblables (LAZRAK et al., 2009 ; SCHALLER et al., 2011). La figure 7.2 illustre cette convergence à travers la mise en évidence de patches de successions d'OCS similaires.

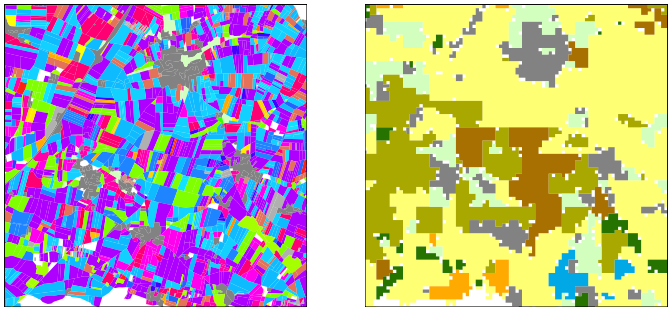


(a)



(b)

FIGURE 7.1 – Illustration de l'application de notre approche de modélisation de l'OTAA sur des données de télédétection (cas du Bassin versant du Yar, Bretagne, adaptée de (LAZRAK et al., Article en cours de rédaction)). (a) Données d'entrée sous forme d'une série temporelle de cartes d'OCS issues d'images satellitaires. (b) Exemple d'un résultat de segmentation temporo-spatiale à partir de ces données d'entrée.



(a) (b)

FIGURE 7.2 – L'organisation de l'OTAA de manière convergente permet de représenter la mosaïque agricole (a) par une carte de zones agricoles homogènes (b). (Illustration adaptée de LAZRAK et al. (2009))

Cette convergence peut s'expliquer par une prise en charge semblable de contraintes et opportunités territoriales similaires par des agriculteurs qui ont des référentiels techniques proches. C'est ainsi qu'en conformité avec cette hypothèse de convergence, nous avons retrouvé, à l'échelle régionale du territoire agricole, des régularités cohérentes avec des règles de décisions d'assolements identifiées à l'échelle d'un petit nombre d'EA (LAZRAK et al., 2011 ; SCHALLER et al., 2011).

7.2 Apports, limites et perspectives d'amélioration des méthodes proposées

7.2.1 Modélisation temporo-spatiale de l'OTAA

L'approche temporo-spatiale tient compte des spécificités de l'activité agricole (*cf.* section 7.1.2 page 158). Elle permet d'identifier les évolutions temporelles puis de les spatialiser sous forme d'une représentation synthétique et simplifiée de l'organisation de l'activité agricole au sein de territoires de dimensions régionales. Nous avons montré que notre méthode permet d'extraire des régularités temporelles et spatiales qui aident à la description et à la compréhension de l'OTAA.

L'implémentation de cette approche dans la boîte à outils ARPENTAGE, nous a permis de tester sa généricité pour des corpus de données provenant de sources variées (relevés de terrain et images de télédétection). Cette informatisation de la méthode permet ainsi un partitionnement automatisable de l'espace agricole en unités paysagères qui rappellent les d'unités agro-physionomiques (UAP) dont l'élaboration était jusque là liée à du savoir-faire d'experts qui devient rare à trouver (THINON et DEFFONTAINES, 1999 ; DEFFONTAINES et THINON, 2001 ; THINON, 2003 ; JACOPIN, 2011). Afin d'évaluer la similitude entre les cartes obtenues par ces deux approches, une étude comparative devrait être poursuivie dans un territoire agricole ayant fait l'objet de l'élaboration d'une carte d'UAP et disposant d'une base de données spatio-temporelles d'OCS. Cette similitude trouve sa raison d'être dans le partage de ces deux méthodes d'une base théorique commune : les Unités Agronomiques (UA) qui sont des portions de territoires de

taille nettement supérieure à celle des parcelles agricoles, et qui présentent une relative égale organisation des systèmes de cultures.

En revanche, notre méthode de modélisation de l'OTAA ne modélise que les polygones de la mosaïque parcellaire. Les éléments paysagers linéaires ou ponctuels comme les réseaux bocagers en bordures de champs ou les arbres et arbustes isolés au sein de l'exploitation agricole ne sont pas pris en compte malgré leur rôle potentiellement important pour l'environnement et la biodiversité (BAUDRY et al., 1998 ; THENAIL, 2002 ; LE COEUR et al., 2002). Un couplage de notre approche avec des approches tenant compte de ces entités spatiales comme celle développée par LE BER et al. (2012) serait une piste à explorer en vue d'enrichir la modélisation de l'OTAA.

Enfin, cette méthode pourrait être étendue à la modélisation de l'organisation territoriale des itinéraires techniques. Cette perspective d'amélioration se résume à la question suivante : pourrions nous rendre intelligible la succession des opérations techniques composant l'itinéraire technique comme nous avons rendu intelligible la succession des couverts végétaux composant les successions de cultures ? Cette perspective nous semble intéressante et opérante avec notre méthode. Nous l'envisagerions en blocs. Par exemple pour le bloc « travail du sol » nous pourrions fouiller les successions temporelles des opérations du travail du sol, identifier des régularités puis, les spatialiser. Idem pour d'autres blocs comme la protection de cultures.

7.2.2 Couplage des régularités stochastiques à l'échelle régionale avec des règles de décision d'assolement à l'échelle de l'exploitation agricole

Les travaux cherchant à expliquer l'organisation et les dynamiques d'usage des sols peuvent être classés dans l'une de deux catégories d'approches (OVERMARS et al., 2007) :

- Une catégorie d'approches **inductives** (appelées aussi « **from pattern to process** »). Les travaux appartenant à cette catégorie partent d'une certaine connaissance de l'organisation du territoire d'étude et d'une liste de facteurs potentiellement explicatifs, et cherchent, par inférence statistique, des corrélations entre des patrons d'organisation et ces facteurs explicatifs. Toutefois, ces corrélations ne peuvent être interprétées comme des liens de causalité et peinent à construire une théorie cohérente.
- Une catégorie d'approches **déductives** (appelées aussi « **from process to pattern** »). Les travaux appartenant à cette catégorie partent d'une théorie et la testent en faisant des simulations de scénarios d'organisation. Une limite majeure de ces approches est l'absence de théories de localisations robustes.

Nous avons proposé dans cette thèse, en étroite collaboration avec la thèse de SCHALLER (2011), un cadre méthodologique visant à construire des bases théoriques pour comprendre l'OTAA et ses dynamiques. Nous avons adopté un point de vue intégrateur fondé sur les décisions d'assolements mises en œuvre par les agriculteurs, identifiées par des enquêtes à l'échelle de l'EA, et vérifiées à l'échelle régionale du territoire agricole (LAZRAK et al., 2011 ; SCHALLER et al., 2011). Ce cadre méthodologique fait

dialoguer des régularités stochastiques identifiées à l'échelle régionale avec des règles de décisions d'assolements identifiées à l'échelle de l'EA. Les résultats obtenus suggèrent une bonne complémentarité entre ces deux niveaux d'échelle. La méthode mise en œuvre dans le cadre de cette thèse a permis d'explorer un premier niveau d'intelligibilité en proposant une première explication des régularités identifiées, et en vérifiant, à l'échelle régionale, la généricité des règles de décisions identifiées à l'échelle de l'EA. Des allers-retours itératifs entre règles et régularités permettraient d'explorer des niveaux croissants d'intelligibilité de l'OTAA : la fouille de données révèlerait des régularités dont la recherche de l'interprétation guiderait la recherche de règles de décisions d'assolements au niveau de l'EA. Ces règles de décisions pourraient être à leur tour vérifiées à l'échelle régionale du territoire agricole en vue d'en définir un domaine de validité. Ce cadre méthodologique permettrait, à terme, d'adopter une approche déductive comme préconisée par OVERMARS et al. (2007), mais seulement après avoir construit une théorie de localisation solide constituée d'un corpus de règles de décisions d'assolements valides. Ce système à base de règles permettrait de réaliser des simulations de réorganisation des EA et aiderait à construire, par agrégation, des mosaïques paysagères agricoles compatibles avec la conservation des ressources naturelles.

Prenons un exemple pour illustrer cette démarche d'accroissement de l'intelligibilité. Deux points d'entrées sont possibles :

- par une régularité identifiée par fouille de données que l'on souhaite interpréter ; ou
- par une règle de décision d'assolement identifiée à l'échelle d'un petit échantillon EA, que l'on souhaite vérifier à l'échelle du territoire agricole.

Nous avons par exemple constaté la régularité statistique suivante :

> « Le Maïs est devenu plus fréquent à proximité des prairies et moins fréquent à proximité des cultures commerciales (tounesol, blé, colza). »

L'expert agronome (SCHALLER et al., 2011) a tenté d'expliquer cette régularité en se basant sur ses résultats d'enquêtes :

> « Quand les agriculteurs ont accès à des sols profonds sur leur territoire d'exploitation, le maïs est cultivé en monoculture au voisinage des prairies situées traditionnellement dans les sols profonds et humides. Lorsque les agriculteurs n'ont pas accès à des sols profonds, leurs choix de cultures dépendent en grande partie du système d'exploitation, des besoins fourragers annuels et de la capacité d'irrigation (modulée par un risque de restriction). Par conséquent, le maïs a tendance à diminuer dans les exploitations de grandes cultures où il n'est pas une culture prioritaire, alors qu'il est maintenu dans les exploitations d'élevage avec accès à l'irrigation. En parallèle, la production du maïs irrigué étant devenue de plus en plus risquée, les éleveurs ont tendance à étendre les prairies pour compléter la production du fourrage. Le maïs est donc essentiellement maintenu sur les exploitations d'élevage avec des zones de prairies importantes sur leur territoire agricole. » (Extrait provenant de la section 3.4.1 page 69, sur la base de SCHALLER et al. (2011)).

Cette explication a valeur d'une hypothèse non validée mais seulement établie par les pratiques d'un faible échantillon d'agriculteurs enquêtés. Passer à un deuxième niveau d'intelligibilité consisterait à décomposer l'explication proposée par l'expert agronome en règles vérifiables par fouille de données, et à vérifier ces règles dans l'ensemble du territoire d'étude, nous aurions quelque chose du genre :

- **Maïs maintenu** dans les exploitations possédant des **sols profonds**
- **Maïs maintenu** dans les exploitations d'**élevage** disposant d'**irrigation**
- etc.

D'autres règles complémentaires devraient également être vérifiées, comme :

- **Maïs diminue** dans les sols **non profonds**
- **Maïs diminue** dans les exploitations de **non élevage**
- **Maïs diminue** dans les exploitations **sans possibilité d'irrigation**
- etc.

La vérification de la cohérence de ces règles d'explications nécessite de disposer d'une carte pédologique qui traduise bien la notion de sol profond des agriculteurs, et d'une typologie des exploitations agricole permettant de distinguer les exploitations d'élevage, ainsi que de l'information sur les surfaces irrigables dans le territoire d'étude. Des fouilles menées de la sorte pourraient permettre de valider complètement ou partiellement les explications du premier niveau et de leur affecter un domaine de validité.

A l'état actuel de la méthode de fouille, toute règle logique n'est pas implémentable en fouille de régularités correspondantes [2]. Une perspective d'amélioration serait de définir un cadre formel de critères auquel les règles doivent répondre afin que leur traduction en modèles de fouille permette d'obtenir des résultats rigoureux.

Cette confrontation entre recherche de régularités stochastiques et établissement de règles logiques permet une rigueur nouvelle dans l'établissement des règles sous forme explicite permettant ensuite une fouille de données et l'établissement des régularités sous expertise préalable (fouille ciblée).

Enfin, notre méthode de recherche de régularités avec les HMM2 s'inscrit plus dans une démarche descriptive qu'explicative. Pour décrire l'OTAA nous nous sommes appuyés sur un paramètre de fouille : l'OCS, auquel nous avons ajouté, lors d'une expérimentation avec les écologues du CEBC [3], un deuxième paramètre : l'occupation (ou non) de la parcelle par un nid de busard. Dans une démarche explicative, des fouilles composites pourraient coupler l'OCS à des facteurs potentiellement explicatifs (comme la géologie, la pédologie, la distance — ou l'accessibilité — à certaines infrastructures, ...) en vue de retrouver les n-uplets les plus fréquents et leurs dépendances fonctionnelles qui traduiraient les contraintes existant entre les variables explicatives et la variable à expliquer. Ces dépendances pourraient ensuite être confrontées, selon le même schéma méthodologique présenté dans cette section, avec les règles identifiées par enquêtes au niveau des exploitations agricoles.

2. Voir par exemple le cas de la conduite de la luzerne en petites bandes contiguës avec des âges différents afin de garder une même qualité nutritive du fourrage et d'en faciliter la gestion. La généricité de cette règle n'a pu être vérifiée, et nous n'avons pas pu trancher de sa non généricité. *Cf.* SCHALLER et al., 2011, pages 442-444.

3. Centre d'Études Biologiques de Chizé.

7.3 Pour conclure : contribution à l'agronomie numérique des territoires

Cette thèse est une contribution à l'intelligence artificielle, et plus particulièrement à la fouille de données pour laquelle nous avons proposé un nouveau cadre de travail adapté à la recherche et à la pluridisciplinarité (*cf.* section 7.3.1). C'est aussi une contribution à l'agronomie, et plus particulièrement celle des territoires (BENOÎT et al., 2007) car elle tend à comprendre le mode selon lequel les agriculteurs organisent dans le temps et l'espace l'allocation de leurs parcelles compte tenu des contraintes et des opportunités. En agronomie des territoires notre contribution se présente sous forme du développement d'une méthode de description de l'OTAA et par la proposition d'un cadre méthodologique permettant d'interagir avec d'autres questions traitées par cette discipline (*cf.* section 7.3.2 page 167).

7.3.1 Un nouveau cadre de fouille de données adapté à la recherche et à la pluridisciplinarité

Le cadre classique de réalisation de la fouille de données fait intervenir une seule compétence : un expert du domaine relatif aux données appelé aussi l'analyste (BENDAOUD et al., 2005 ; NAPOLI, 2005), qui est chargé de :

- diriger le processus de fouille de données
- et d'interpréter les éléments d'information extraits pour les traduire en connaissances.

Ce cadre ainsi défini nécessitait d'être adapté pour convenir à la nature de notre approche caractérisée par plusieurs particularités :

- à la différence du cadre classique de fouille de données où l'outil de fouille est au point, notre outil de fouille « ARPENTAGE » se développait en fonction des besoins rencontrés ;
- les liens étroits entre la base de données spatio-temporelle d'OCS et leur contexte historique et territorial, nécessitent de l'expert du domaine une connaissance suffisante du territoire agricole étudié. L'expert, un agronome de terrain, qui acquiert des connaissances par ses interactions avec les acteurs agissant sur l'organisation de l'activité agricole dans le territoire d'étude ;
- enfin, la recherche de liens entre l'OTAA et les questions environnementales (écologie, hydrologie), nécessite la collaboration des spécialistes de ces domaines avec l'expert agronome qui possède une connaissance du territoire d'étude, et l'analyste [4] qui dirige la fouille de données.

Ces particularités concernent 3 étapes de travail inter-connectées :

1. Le développement de l'outil de fouille de données. Cette étape a été menée par Jean-François Mari (informaticien à l'équipe d'extraction et représentation de connaissances « Orpailleur » du Loria).
2. La fouille de données au sens strict. Cette étape a été assurée par moi-même et consistait à préparer les corpus de données, et à définir les modèles de fouille en fonction des besoins des experts des

4. À la différence du cadre classique de fouille de données, nous distinguons l'analyste de l'expert du domaine auxquels nous affectons des rôles différents.

domaines (agronomie, écologie, hydrologie), et à présenter les résultats de fouille sous des formes visuelles « parlantes » pour l'expert du domaine.

3. L'interprétation des résultats de fouille. Cette étape a été prise en charge par deux types d'experts :
   – une agronome de terrain (Noémie SCHALLER qui était doctorante dans le cadre du projet « BiodivAgriM ») ;
   – et par des écoloques au CEBC spécialistes en écologie comportementale et conservation des espèces menacées.

Notre cadre de fouille de données a été orienté vers un objectif de mise au point d'une méthode générique de fouille et de modélisation de l'OTAA. La mise au point est effectivement terminée. Cependant cette méthode requiert, quant à son utilisation, la collaboration de certains intervenants à savoir :

– l'analyste qui dirige la fouille de données ;
– les experts des domaines relatifs aux données, au territoire d'étude, et à la problématique environnementale.

Ces intervenants définissent les besoins de fouille et interprètent les résultats. C'est une collaboration interactive et itérative.

7.3.2 Contribution à l'agronomie des territoires

L'agronomie des territoires cherche à comprendre les liens entre systèmes agricoles et ressources naturelles à l'échelle des territoires agricoles. Elle articule 3 points :

i. L'activité agricole à l'échelle de l'exploitation et à l'échelle régionale, ce premier point cherche à répondre à la question : quelle OTAA ?
ii. L'impact de cette OTAA sur les ressources naturelles et cherche à répondre à la question : quels sont les résultats de l'OTAA ?
iii. Les raisons de l'organisation de l'OTAA exprimées par les agriculteurs, ce point cherche à répondre à la question : pourquoi cette OTAA ?

L'agronomie des territoires utilise des outils de l'agronomie et de la géoagronomie, et cherche à développer ses propres outils. Ma thèse propose des méthodes utiles pour l'agronomie des territoires permettant d'explorer le point (i), et d'interagir avec les points (ii) et (iii). La mobilisation des outils informatiques dans cet effort de modélisation est ainsi une contribution à l'agronomie numérique des territoires.

7.4 Références

BAUDRY, J, A JOUIN et C THENAIL (1998). « La diversité des bordures de champ dans les exploitations agricoles de pays de bocage ». Dans : *Etudes et Recherches sur les Systèmes Agraires* 31, p. 117–134.

BAUM, LE, T PETRIE, G SOULES et N WEISS (1970). « A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains ». Dans : *The annals of mathematical statistics* 41.1, p. 164–171.

BENDAOUD, R, Y TOUSSAINT et A NAPOLI (2005). « Hiérarchisation des règles d'association en fouille de textes ». Français. Dans : *Revue des Sciences et Technologies de l'Information (Série Ingénierie des Systèmes d'Information)* 1, p. 263–274. URL : http://hal.inria.fr/inria-00000436.

BENMILOUD, B et W PIECZYNSKI (1995). « Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images ». Dans : *Traitement du signal* 12.5, p. 433–454.

BENOÎT, M (2006). « Organisation territoriale des activités agricoles ». Dans : *Acteurs et territoires locaux: vers une géoagronomie de l'aménagement*. Editions Quae, p. 87–89.

BENOÎT, M, C MIGNOLET, S HERRMANN, D RIZZO, C MOONEN, P BARBERI, M GALLI, E BONARI, N SILVERSTRI, C THENAIL, S LARDON, H RAPEY, E MARRACCINI, F LE BER et JM MEYNARD (2007). « Landscape designed by farming systems: a challenge for landscape agronomists in Europe ». Dans : *Farming Systems Design 2007, Methodologies for Integrated Analysis of Farm Production Systems*. Catania, Sicilia, Italy, p. 2.

BONTHOUX, S (2011). « Relations spatiales et temporelles entre les communautés d'oiseaux et les paysages agricoles ». Thèse de doct. Institut National Polytechnique de Toulouse.

CAPLIER, PA (1995). « Modèle markovien de détection de mouvement dans les séquences d'images: approche spatio-temporelle et mises en œuvre temps réel ». Thèse de doct. Institut national polytechnique de Grenoble, Grenoble, France.

CASTELLAZZI, MS, GA WOOD, PJ BURGESS, J MORRIS, KF CONRAD et JN PERRY (2008). « A systematic representation of crop rotations ». Dans : *Agricultural Systems* 97.1-2, p. 26–33.

CREWS-MEYER, KA (2004). « Agricultural landscape change and stability in northeast Thailand: historical patch-level analysis ». Dans : *Agriculture, ecosystems & environment* 101.2-3, p. 155–169.

DEFFONTAINES, JP, E LANDAIS et P PIERRET (1997). « Le temps long et le temps rond des paysages agricoles ». Dans : *Programme environnement, vie et sociétés. Journées*, p. 71–81.

DEFFONTAINES, JP, C THENAIL et J BAUDRY (1995). « Agricultural systems and landscape patterns: how can we build a relationship? » Dans : *Landscape and urban planning* 31.1-3, p. 3–10.

DEFFONTAINES, JP et P THINON (2001). « Des entités spatiales significatives pour l'activité agricole et pour les enjeux environnementaux et paysagers. Contribution à une agronomie du territoire ». Dans : *Courrier de l'Environnement, Inra* 44, p. 13–28.

GLENN-LEWIN, DC et E VAN DER MAAREL (1992). « Plant Succession: Theory and Prediction ». Dans : sous la dir. de RK PEET, DC GLENN-LEWIN et TT VEBLEN. Chapman & Hall. Chap. Patterns and processes of vegetation dynamics, p. 11–44.

JACOPIN, R (2011). « Paysages et pratiques des agriculteurs dans le Sud du Plateau Lorrain : logiques d'organisation et effect sur l'environnement. » Thèse de doct. Université Nancy 2.

JELINEK, F (1976). « Continuous speech recognition by statistical methods ». Dans : *Proceedings of the IEEE* 64.4, p. 532–556.

LADET, S, M DECONCHAT, C MONTEIL, J LACOMBE et G BALENT (2005). « chaîne de Markov, automate cellulaire, évaluation multicritère, SIG, changement d'occupation du sol ». Dans : *Revue Internationale de Géomatique* 15.2, p. 159–173.

LAZRAK, EG, M BENOÎT et J-F MARI (Article en cours de rédaction). « Segmentation temporo-spatiale du bassin versant du Yar fondée sur des occupations du sol télédétectées ».

LAZRAK, EG, J-F MARI et M BENOÎT (2009). « Landscape regularity modelling for environmental challenges in agriculture ». Dans : *Landscape Ecology* 25.2, p. 169–183.

LAZRAK, EG, N SCHALLER et J-F MARI (2011). « Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol ». Français. Dans : *Atelier en marge d'EGC 2011*. Brest, France.

LE BER, F, C LAVIGNE et S DA SILVA (2012). « Structure analysis of hedgerows and other perennial landscape lines in two French agricultural landscapes ». Anglais. Dans : *Proceedings of 15th AGILE International Conference*. Avignon, France, p. 6.

LE COEUR, D, J BAUDRY, F BUREL et C THENAIL (2002). « Why and how we should study field boundary biodiversity in an agrarian landscape context ». Dans : *Agriculture, ecosystems & environment* 89.1-2, p. 23–40.

MARTIN, P, A JOANNON, C MIGNOLET, V SOUCHÈRE et C THENAIL (2006). « Systèmes de culture et territoires: cas des questions environnementales ». Dans : *L'agronomie aujourd'hui*. Versailles : Quae éditions, p. 253–283.

MIGNOLET, C, C SCHOTT et M BENOÎT (2007). « Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale ». Dans : *Science of the Total Environment* 375.1-3, p. 13–32.

MOTTET, A, S LADET, N COQUE et A GIBON (2006). « Agricultural land-use change and its drivers in mountain landscapes: A case study in the Pyrenees ». Dans : *Agriculture, ecosystems & environment* 114.2-4, p. 296–310.

NAPOLI, A (2005). « Chapter 41 - A Smooth Introduction to Symbolic Methods for Knowledge Discovery ». Dans : *Handbook of Categorization in Cognitive Science*. Sous la dir. d'Henri COHEN et Claire LEFEBVRE. Oxford : Elsevier Science Ltd, p. 913 –933. URL : http://www.sciencedirect.com/science/article/pii/B9780080446127500962.

OVERMARS, KP, WT DE GROOT et MGA HUIGEN (2007). « Comparing Inductive and Deductive Modeling of Land Use Decisions: Principles, a Model and an Illustration from the Philippines ». Dans : *Human Ecology* 35 (4), p. 439–452. URL : http://dx.doi.org/10.1007/s10745-006-9101-6.

PAPY, F et A TORRE (2002). « Quelles organisations territoriales pour concilier production agricole et gestion des ressources naturelles ». Dans : *Etudes et Recherches sur les Systèmes Agraires et le Développement* 33, p. 151–170.

POUDEVIGNE, I et D ALARD (1997). « Landscape and agricultural patterns in rural areas: a case study in the Brionne Basin, Normandy, France ». Dans : *Journal of Environmental Management* 50.4, p. 335–349.

SCHALLER, N (2011). « Modélisation des décisions d'assolement des agriculteurs et de l'organisation spatiale des cultures dans les territoires de polyculture-élevage ». Thèse de doct. L'Institut des Sciences et Industries du Vivant et de l'Environnement (AgroParisTech).

SCHALLER, N, EG LAZRAK, P MARTIN, J-F MARI, C AUBRY et M BENOÎT (2011). « Combining farmers' decision rules and landscape stochastic regularities for landscape modelling ». Anglais. Dans : *Landscape Ecology*.

SCHIETTECATTE, W, L D'HONDT, WM CORNELIS, ML ACOSTA, Z LEAL, N LAUWERS, Y ALMOZA, GR ALONSO, J DÍAZ, M RUÍZ et D GABRIELS (2008). « Influence of landuse on soil erosion risk in the Cuyaguateje watershed (Cuba) ». Dans : *CATENA* 74.1, p. 1–12.

SEBILLOTTE, M (1974). « Agronomie et agriculture. Essai d'analyse des tâches de l'agronome ». Dans : *Cahiers ORSTOM, série Biologie* 24, p. 3–25.

— (1990). « Système de culture, un concept opératoire pour les agronomes ». Dans : *Les systèmes de culture*. Un Point sur. L Combe, p. 165–196.

SIMON, G (2012). « Modélisations multi-scalaires des dynamiques urbaines dans la longue durée: l'exemple du quartier abbatial de Vendôme (41) ». Dans : *Cybergeo: European Journal of Geography*.

SIMPSON, JW, REJ BOERNER, MN DE MERS, LA BERNS, FJ ARTIGAS et A SILVA (1994). « Forty-eight years of landscape change on two contiguous Ohio landscapes ». Dans : *Landscape Ecology* 9.4, p. 261–270.

THENAIL, C (2002). « Relationships between farm characteristics and the variation of the density of hedgerows at the level of a micro-region of bocage landscape. Study case in Brittany, France ». Dans : *Agricultural Systems* 71.3, p. 207–230.

THENAIL, C, A JOANNON, M CAPITAINE, V SOUCHÈRE, C MIGNOLET, N SCHERMANN, F DI PIETRO, Y PONS, C GAUCHEREL, V VIAUD et al. (2009). « The contribution of crop-rotation organization in farms to crop-mosaic patterning at local landscape scales ». Dans : *Agriculture, Ecosystems & Environment* 131.3-4, p. 207–219.

THINON, P (2003). « Les unités agro-physionomiques: quels usages? Quelle prise en compte du temps? » Dans : *Actes du colloque international*. Montpellier, France : P Dugué et Ph Jouve.

THINON, P et JP DEFFONTAINES (1999). « Partage de l'espace rural pour la gestion de problèmes environnementaux et paysagers dans le Vexin français ». Dans : *Cah Agr* 8, p. 373–87.

USHER, MB (1992). « Plant Succession: Theory and Prediction ». Dans : sous la dir. de RK PEET, DC GLENN-LEWIN et TT VEBLEN. Chapman & Hall. Chap. Statistical models of succession, p. 215–246.

WELCH, LR (2003). « Hidden markov models and the baum-welch algorithm ». Dans : *IEEE Information Theory Society Newsletter* 53.4, p. 1–10.

Références générales

ABRIAL, D, L AZIZI, M CHARRAS-GARRIDO et F FORBES (2010). « Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie à l'aide de champs de Markov cachés ». Dans : *42èmes Journées de Statistique*. Marseille, France, France. URL : http://hal.inria.fr/inria-00494838.

AVIRON, S, P KINDLMANN et F BUREL (2007). « Conservation of butterfly populations in dynamic landscapes: The role of farming practices and landscape mosaic ». Dans : *ecological modelling* 205.1-2, p. 135–145.

AZIZI, L, F FORBES, S DOYLE, M CHARRAS-GARRIDO et D ABRIAL (2011). *Spatial risk mapping for rare disease with hidden Markov fields and variational EM*. Anglais. Rapport de recherche RR-7572. INRIA.

BAUDRY, J, A JOUIN et C THENAIL (1998). « La diversité des bordures de champ dans les exploitations agricoles de pays de bocage ». Dans : *Etudes et Recherches sur les Systèmes Agraires* 31, p. 117–134.

BAUDRY, J, F BUREL, S AVIRON, M MARTIN, A OUIN, G PAIN et C THENAIL (2003). « Temporal variability of connectivity in agricultural landscapes: do farming activities help? » Dans : *Landscape ecology* 18.3, p. 303–314.

BAUM, LE, T PETRIE, G SOULES et N WEISS (1970). « A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains ». Dans : *The annals of mathematical statistics* 41.1, p. 164–171.

BENDAOUD, R, Y TOUSSAINT et A NAPOLI (2005). « Hiérarchisation des règles d'association en fouille de textes ». Français. Dans : *Revue des Sciences et Technologies de l'Information (Série Ingénierie des Systèmes d'Information)* 1, p. 263–274. URL : http://hal.inria.fr/inria-00000436.

BENMILOUD, B et W PIECZYNSKI (1995). « Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images ». Dans : *Traitement du signal* 12.5, p. 433–454.

BENOÎT, M (2006). « Organisation territoriale des activités agricoles ». Dans : *Acteurs et territoires locaux: vers une géoagronomie de l'aménagement*. Editions Quae, p. 87–89.

BENOÎT, M et F PAPY (1998). « La place de l'agronomie dans la problématique environnementale ». Dans : *Dossiers de l'environnement INRA* 17, p. 53–71. URL : http://www.inra.fr/dpenv/benoid17.htm.

BENOÎT, M, C MIGNOLET, S HERRMANN, D RIZZO, C MOONEN, P BARBERI, M GALLI, E BONARI, N SILVERSTRI, C THENAIL, S LARDON, H RAPEY, E MARRACCINI, F LE BER et JM MEYNARD (2007). « Landscape designed by farming systems: a challenge for landscape agronomists in Europe ». Dans : *Farming Systems Design 2007, Methodologies for Integrated Analysis of Farm Production Systems*. Catania, Sicilia, Italy, p. 2.

BENOÎT, M, D RIZZO, E MARRACCINI, AC MOONEN, M GALLI, S LARDON, H RAPEY, C THENAIL et E BONARI (Soumis). « Landscape agronomy : a new perspective for research on agricultural landscapes ». Dans : *Landscape Ecology*.

BERTHOD, M, Z KATO, S YU et J ZERUBIA (1996). « Bayesian image classification using Markov random fields ». Dans : *Image and Vision Computing* 14.4, p. 285–295.

BONNEMAIRE, J (1995). *Pays, paysans, paysages: dans les Vosges du Sud : les pratiques agricoles et la transformation de l'espace*. Institut national de la recherche agronomique.

BONTHOUX, S (2011). « Relations spatiales et temporelles entre les communautés d'oiseaux et les paysages agricoles ». Thèse de doct. Institut National Polytechnique de Toulouse.

BÜHLER, E-A, A CAMARA, S LOPEZ-RIDAURA et C-T SOULARD (2010). « Farms and territories: crossing agronomy and geography to elaborate multifunctional farming systems ». Français. Dans : *Innovation and Sustainable Development in Agriculture and Food*. Sous la dir. d'E COUDEL, H DEVAUTOUR, C-T SOULARD et B HUBERT. Montpellier, France : Cirad-Inra-SupAgro, p. 16.

BUREL, F et J BAUDRY (2005). « Habitat quality and connectivity in agricultural landscapes: The role of land use systems at various scales in time ». Dans : *Ecological Indicators* 5.4, p. 305–313.

CAPLIER, PA (1995). « Modèle markovien de détection de mouvement dans les séquences d'images: approche spatio-temporelle et mises en œuvre temps réel ». Thèse de doct. Institut national polytechnique de Grenoble, Grenoble, France.

CASTELLAZZI, MS, GA WOOD, PJ BURGESS, J MORRIS, KF CONRAD et JN PERRY (2008). « A systematic representation of crop rotations ». Dans : *Agricultural Systems* 97.1-2, p. 26–33.

CELEUX, G et G GOVAERT (1992). « A classification EM algorithm for clustering and two stochastic versions ». Dans : *Computational Statistics & Data Analysis* 14.3, p. 315–332.

COPPEDGE, BR, DM ENGLE, RE MASTERS et MS GREGORY (2001). « Avian response to landscape change in fragmented southern Great Plains grasslands ». Dans : *Ecological Applications* 11.1, p. 47–59.

CORGNE, S (2004). « Hiérarchisation des facteurs de changements de l'occupation hivernale des sols: Application au bassin versant du Yar (Bretagne) ». Dans : *Norois* 193, p. 17–29.

CREWS-MEYER, KA (2004). « Agricultural landscape change and stability in northeast Thailand: historical patch-level analysis ». Dans : *Agriculture, ecosystems & environment* 101.2-3, p. 155–169.

DEFFONTAINES, JP, E LANDAIS et P PIERRET (1997). « Le temps long et le temps rond des paysages agricoles ». Dans : *Programme environnement, vie et sociétés. Journées*, p. 71–81.

DEFFONTAINES, JP, C THENAIL et J BAUDRY (1995). « Agricultural systems and landscape patterns: how can we build a relationship? » Dans : *Landscape and urban planning* 31.1-3, p. 3–10.

DEFFONTAINES, JP et P THINON (2001). « Des entités spatiales significatives pour l'activité agricole et pour les enjeux environnementaux et paysagers. Contribution à une agronomie du territoire ». Dans : *Courrier de l'Environnement, Inra* 44, p. 13–28.

DOGLIOTTI, S, WAH ROSSING et MK VAN ITTERSUM (2003). « ROTAT, a tool for systematically generating crop rotations ». Dans : *European Journal of Agronomy* 19.2, p. 239–250.

DONALD, PF, RE GREEN et MF HEATH (2001). « Agricultural intensification and the collapse of Europe's farmland bird populations ». Dans : *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268.1462, p. 25.

ENG, C, C ASTHANA, B AIGLE, S HERGALANT, J-F MARI et P LEBLOND (2009). « A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods ». Dans : *Journal of Computational Biology* 16.9, p. 1211–1225.

FAOSTAT (2009). http://faostat.fao.org/site/377/default.aspx.

FINE, S, Y SINGER et N TISHBY (1998). « The Hierarchical Hidden Markov Model: Analysis and Applications ». Dans : *Machine Learning* 32, p. 41–62.

GLENN-LEWIN, DC et E VAN DER MAAREL (1992). « Plant Succession: Theory and Prediction ». Dans : sous la dir. de RK PEET, DC GLENN-LEWIN et TT VEBLEN. Chapman & Hall. Chap. Patterns and processes of vegetation dynamics, p. 11–44.

GLP (2005). *Science Plan and Implementation Strategy*. Rap. tech. IGBP Report No. 53/IHDP Report No. 19. IGBP Secretariat, Stockholm, p. 64.

HÉNIN, D (1999). *De la méthode en agronomie*. Ecologie et agronomie appliquées. l'Harmattan.

HÉNIN, S et JP DEFFONTAINES (1970). « Principe et utilité de l'étude des potentialités agricoles régionales ». Dans : *Comptes Rendus de l'Académie d'Agriculture de France*, p. 463–472.

HERRERA, JM et M LIEDGENS (2009). « Leaching and utilization of nitrogen during a spring wheat catch crop succession ». Dans : *Journal of environmental quality* 38.4, p. 1410–1419.

HOUET, T (2006). « Occupation des sols et gestion de l'eau: modélisation prospective en paysage agricole fragmenté (Application au SAGE du Blavet) ». Thèse de doct. Université Rennes 2.

INCHAUSTI, P et V BRETAGNOLLE (2005). « Predicting short-term extinction risk for the declining Little Bustard Tetrax tetrax in intensive agricultural habitats ». Dans : *Biological conservation* 122.3, p. 375–384.

INTERNATIONAL DE LA LANGUE FRANÇAISE, Conseil (1999). *Dictionnaire d'agriculture: français-anglais-allemand*. Le Conseil.

JACOPIN, R (2011). « Paysages et pratiques des agriculteurs dans le Sud du Plateau Lorrain : logiques d'organisation et effect sur l'environnement. » Thèse de doct. Université Nancy 2.

JELINEK, F (1976). « Continuous speech recognition by statistical methods ». Dans : *Proceedings of the IEEE* 64.4, p. 532–556.

KROGH, A (1998). « Chapter 4 An introduction to hidden Markov models for biological sequences ». Dans : *Computational Methods in Molecular Biology*. Sous la dir. de SL SALZBERG, DB SEARLS et S KASIF. T. 32. New Comprehensive Biochemistry. Elsevier, p. 45 –63.

LACHIVER, M (1997). *Dictionnaire du monde rural: les mots du passé*. Fayard.

LADET, S, M DECONCHAT, C MONTEIL, J LACOMBE et G BALENT (2005). « chaîne de Markov, automate cellulaire, évaluation multicritère, SIG, changement d'occupation du sol ». Dans : *Revue Internationale de Géomatique* 15.2, p. 159–173.

LANDAIS, E et JP DEFFONTAINES (1990). « Comprendre la gestion d'un espace pastoral. Étude monographique des pratiques d'un berger d'estive dans les Alpes du Sud ». Dans : *Recherches sur les Systèmes Herbagers*, p. 189–197.

LAZRAK, EG, M BENOÎT et J-F MARI (2009). « Fouille de données à l'aide de modèles stochastiques: segmentation temporo-spatiale des successsions de cultures d'un territoire agricole à l'aide de HMM2 ». Dans : *STIC 2009 Environnement*. Calais, France.

— (2010). « Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales ». Dans : *International Conference on Integrative Landscape Modelling*. Montpellier, France.

LAZRAK, EG, M BENOÎT et J-F MARI (Article en cours de rédaction). « Segmentation temporo-spatiale du bassin versant du Yar fondée sur des occupations du sol télédétectées ».

LAZRAK, EG, J-F MARI et M BENOÎT (2009). « Landscape regularity modelling for environmental challenges in agriculture ». Dans : *Landscape Ecology* 25.2, p. 169–183.

LAZRAK, EG, N SCHALLER et J-F MARI (2011). « Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol ». Français. Dans : *Atelier en marge d'EGC 2011*. Brest, France.

LE BER, F, C LAVIGNE et S DA SILVA (2012). « Structure analysis of hedgerows and other perennial landscape lines in two French agricultural landscapes ». Anglais. Dans : *Proceedings of 15th AGILE International Conference*. Avignon, France, p. 6.

LE BER, F, M BENOÎT, C SCHOTT, J-F MARI et C MIGNOLET (2006). « Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software ». Dans : *Ecological Modelling* 191.1, p. 170–185.

LE COEUR, D, J BAUDRY, F BUREL et C THENAIL (2002). « Why and how we should study field boundary biodiversity in an agrarian landscape context ». Dans : *Agriculture, ecosystems & environment* 89.1-2, p. 23–40.

LUNT, ID et PG SPOONER (2005). « Using historical ecology to understand patterns of biodiversity in fragmented agricultural landscapes ». Dans : *Journal of Biogeography* 32.11, p. 1859–1873.

MARI, J-F, J-P HATON et A KRIOUILE (1997). « Automatic Word Recognition Based on Second-Order Hidden Markov Models ». Dans : *IEEE Transactions on Speech and Audio Processing* 5, p. 22–25.

MARI, J-F, EG LAZRAK et M BENOÎT (2010). « Fouille de paysages agricoles: analyse des voisinages des successions d'occupation du sol ». Dans : *Colloque RTE (Raisonnement sur le Temps et l'Espace) en marge de RFIA 2010*. Sous la dir. de F LE BER, G LIGOZAT, O PAPINI et M BOUZID. Caen, France.

MARI, J-F et F LE BER (2006). « Temporal and Spatial Data Mining with Second-Order Hidden Markov Models ». Dans : *Soft Computing*. ISSN:1432-7643 10.5, p. 406–414.

MARI, J-F, F LE BER et M BENOÎT (1998). « Reconnaissance de successions culturales par modèles de Markov : une étude préliminaire ». Dans : *Journées Cassini*. Marne-la-Vallée.

— (1999). « Classification de successions culturales par modèles de Markov ». Dans : *Septième journées de la Société Francophone de Classification - SFC'99*. Colloque avec actes et comité de lecture, p. 177–184.

MARTIN, P, F PAPY, V SOUCHÈRE et A CAPILLON (1998). « Maîtrise du ruissellement et modélisation des pratiques de production ». Dans : *Cahiers Agricultures* 7.2, p. 111–119.

MARTIN, P, A JOANNON, C MIGNOLET, V SOUCHÈRE et C THENAIL (2006). « Systèmes de culture et territoires: cas des questions environnementales ». Dans : *L'agronomie aujourd'hui*. Versailles : Quae éditions, p. 253–283.

MÉROT, A, C AUBRY, M BARBIER, A JOANNON, P MARTIN, C THENAIL et M BENOÎT (Soumis). « Interfacing landscape and agrosystems research at various scales to deal with natural resource preservation : a review ». Dans : *Agriculture Ecosystems & Environment*.

MIGNOLET, C (2008). « Modélisation de l'organisation spatiale des systèmes agricoles et de son évolution dans des démarches d'appui au développement ». Thèse de doct. Institut des sciences et industries du vivant et de l'environnement (Agro Paris Tech).

MIGNOLET, C, C SCHOTT et M BENOÎT (2004). « Spatial dynamics of agricultural practices on a basin territory: a retrospective study to implement models simulating nitrate flow. The case of the Seine basin ». Dans : *Agronomie* 24.4, p. 219–236.

— (2007). « Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale ». Dans : *Science of the Total Environment* 375.1-3, p. 13–32.

MOTTET, A, S LADET, N COQUE et A GIBON (2006). « Agricultural land-use change and its drivers in mountain landscapes: A case study in the Pyrenees ». Dans : *Agriculture, ecosystems & environment* 114.2-4, p. 296–310.

NAPOLI, A (2005). « Chapter 41 - A Smooth Introduction to Symbolic Methods for Knowledge Discovery ». Dans : *Handbook of Categorization in Cognitive Science*. Sous la dir. d'Henri COHEN et Claire LEFEBVRE. Oxford : Elsevier Science Ltd, p. 913 –933. URL : http://www.sciencedirect.com/science/article/pii/B9780080446127500962.

ORMEROD, SJ et AR WATKINSON (2000). « Editors' introduction: birds and agriculture ». Dans : *Journal of applied ecology* 37.5, p. 699–705.

OSTY, PL (1978). « L'exploitation vue comme une système: Diffusion de l'innovation et contribution au développement ». Dans : *Bulletin Technique d'Information* 326, p. 43–49.

OVERMARS, KP, WT DE GROOT et MGA HUIGEN (2007). « Comparing Inductive and Deductive Modeling of Land Use Decisions: Principles, a Model and an Illustration from the Philippines ». Dans : *Human Ecology* 35 (4), p. 439–452. URL : http://dx.doi.org/10.1007/s10745-006-9101-6.

PAPY, F et A TORRE (2002). « Quelles organisations territoriales pour concilier production agricole et gestion des ressources naturelles ». Dans : *Etudes et Recherches sur les Systèmes Agraires et le Développement* 33, p. 151–170.

POUDEVIGNE, I et D ALARD (1997). « Landscape and agricultural patterns in rural areas: a case study in the Brionne Basin, Normandy, France ». Dans : *Journal of Environmental Management* 50.4, p. 335–349.

RABINER, LR (1989). « A tutorial on hidden Markov models and selected applications in speech recognition ». Dans : *Proceedings of the IEEE* 77.2, p. 257–286.

RAINI, JA (2009). « Impact of land use changes on water resources and biodiversity of Lake Nakuru catchment basin, Kenya ». Dans : *African Journal of Ecology* 47, p. 39–45.

REY, A (2005). *Dictionnaire culturel en langue française: Coffret en 4 volumes*. Le Robert.

RINDFUSS, RR, SJ WALSH, BL TURNER, J FOX et V MISHRA (2004). « Developing a science of land change: Challenges and methodological issues ». Dans : *Proceedings of the National Academy of Sciences of the United States of America* 101.39, p. 13976–13981.

SCHALLER, N (2011). « Modélisation des décisions d'assolement des agriculteurs et de l'organisation spatiale des cultures dans les territoires de polyculture-élevage ». Thèse de doct. L'Institut des Sciences et Industries du Vivant et de l'Environnement (AgroParisTech).

SCHALLER, N, EG LAZRAK, P MARTIN, J-F MARI, C AUBRY et M BENOÎT (2011). « Combining farmers' decision rules and landscape stochastic regularities for landscape modelling ». Anglais. Dans : *Landscape Ecology*.

SCHIETTECATTE, W, L D'HONDT, WM CORNELIS, ML ACOSTA, Z LEAL, N LAUWERS, Y ALMOZA, GR ALONSO, J DÍAZ, M RUÍZ et D GABRIELS (2008). « Influence of landuse on soil erosion risk in the Cuyaguateje watershed (Cuba) ». Dans : *CATENA* 74.1, p. 1–12.

SEBILLOTTE, M (1974). « Agronomie et agriculture. Essai d'analyse des tâches de l'agronome ». Dans : *Cahiers ORSTOM, série Biologie* 24, p. 3–25.

— (1990). « Système de culture, un concept opératoire pour les agronomes ». Dans : *Les systèmes de culture*. Un Point sur. L Combe, p. 165–196.

— (2005). « Agronomes et territoires : Deuxième édition des Entretiens du Pradel ». Dans : sous la dir. de P PREVOST. Harmattan. Chap. Agronomes et territoires. Les trois métiers des agronomes, p. 479–497.

SEBILLOTTE, M et L-G SOLER (1988). « Le concept de modèle général et la compréhension du comportement de l'agriculteur, CR Acad ». Dans : *Agric. Fr* 74, p. 59–70.

— (1990). « Les processus de décision des agriculteurs ». Dans : *Modélisation systémique et système agraire: décision et organisation*. Sous la dir. de J BROSSIER, B VISSAC et JLL MOIGNE. INRA.

SIMON, G (2012). « Modélisations multi-scalaires des dynamiques urbaines dans la longue durée: l'exemple du quartier abbatial de Vendôme (41) ». Dans : *Cybergeo: European Journal of Geography*.

SIMPSON, JW, REJ BOERNER, MN DE MERS, LA BERNS, FJ ARTIGAS et A SILVA (1994). « Forty-eight years of landscape change on two contiguous Ohio landscapes ». Dans : *Landscape Ecology* 9.4, p. 261–270.

STOMPH, TJ, LO FRESCO et H VAN KEULEN (1994). « Land use system evaluation: Concepts and methodology ». Dans : *Agricultural systems* 44.3, p. 243–255.

THENAIL, C (2002). « Relationships between farm characteristics and the variation of the density of hedgerows at the level of a micro-region of bocage landscape. Study case in Brittany, France ». Dans : *Agricultural Systems* 71.3, p. 207–230.

THENAIL, C, A JOANNON, M CAPITAINE, V SOUCHÈRE, C MIGNOLET, N SCHERMANN, F DI PIETRO, Y PONS, C GAUCHEREL, V VIAUD et al. (2009). « The contribution of crop-rotation organization in farms to crop-mosaic patterning at local landscape scales ». Dans : *Agriculture, Ecosystems & Environment* 131.3-4, p. 207–219.

THINON, P (2003). « Les unités agro-physionomiques: quels usages? Quelle prise en compte du temps? » Dans : *Actes du colloque international*. Montpellier, France : P Dugué et Ph Jouve.

THINON, P et JP DEFFONTAINES (1999). « Partage de l'espace rural pour la gestion de problèmes environnementaux et paysagers dans le Vexin français ». Dans : *Cah Agr* 8, p. 373–87.

TSCHARNTKE, T, AM KLEIN, A KRUESS, I STEFFAN-DEWENTER et C THIES (2005). « Landscape perspectives on agricultural intensification and biodiversity-ecosystem service management ». Dans : *Ecology Letters* 8.8, p. 857–874.

TURNER, BL, EF LAMBIN et A REENBERG (2007). « The emergence of land change science for global environmental change and sustainability ». Dans : *Proceedings of the National Academy of Sciences* 104.52, p. 20666–20671.

TURNER, BL, D SKOLE, S SANDERSON, G FISCHER, L FRESCO et R LEEMANNS (1995). *Land-Use and Land-Cover Change: Science/Research Plan*. Rap. tech., p. 132.

USHER, MB (1992). « Plant Succession: Theory and Prediction ». Dans : sous la dir. de RK PEET, DC GLENN-LEWIN et TT VEBLEN. Chapman & Hall. Chap. Statistical models of succession, p. 215–246.

VERBURG, PH et A VELDKAMP (2001). « The role of spatially explicit models in land-use change research: a case study for cropping patterns in China ». Dans : *Agriculture, ecosystems & environment* 85.1-3, p. 177–190.

VERBURG, PH, K KOK, RG PONTIUS et A VELDKAMP (2006). « Modeling Land-Use and Land-Cover Change ». Dans : *Land-Use and Land-Cover Change*. Sous la dir. d'EF LAMBIN et H GEIST. Global Change – The IGBP Series (closed). Springer Berlin Heidelberg, p. 117–135. ISBN : 978-3-540-32202-3.

VITOUSEK, PM, HA MOONEY, J LUBCHENCO et JM MELILLO (1997). « Human domination of Earth's ecosystems ». Dans : *Science* 277.5325, p. 494.

WELCH, LR (2003). « Hidden markov models and the baum-welch algorithm ». Dans : *IEEE Information Theory Society Newsletter* 53.4, p. 1–10.

Annexe A

Using Markov Models to Mine Temporal and Spatial Data

Cet article est un chapitre du livre : « *New Fundamental Technologies in Data Mining* ». C'est un article méthodologique qui, après une partie théorique commune, rapporte de manière synthétique des travaux récents de fouille de données utilisant des modèles de Markov cachés (HMM) dans divers domaines d'applications. Ma contribution a porté sur l'application des HMM à la fouille des territoires agricoles.

30

Using Markov Models to Mine Temporal and Spatial Data

Jean-François Mari[1], Florence Le Ber[1,2], El Ghali Lazrak[3], Marc Benoît[3]
Catherine Eng[4], Annabelle Thibessard[4] and Pierre Leblond[4]
[1]*LORIA / Inria-Grand Est, Campus scientifique, BP 239, F-54500, Vandœuvre-lès-Nancy*
[2] *ENGEES, 1 Quai Koch, F-67000, Strasbourg*
[3]*INRA, UR 055, SAD-ASTER domaine du Joly, F-88500, Mirecourt*
[4]*Laboratoire de Génétique et de Microbiologie, UHP-INRA, UMR 1128-IFR110, F-54500, Vandœuvre-lès-Nancy*
France

1. Stochastic modelling, temporal and spatial data and graphical models

Markov models represent a powerful way to approach the problem of mining time and spatial signals whose variability is not yet fully understood. Initially developed for pattern matching (Baker, 1974; Geman & Geman, 1984) and information theory (Forney, 1973), they have shown good modelling capabilities in various problems occurring in different areas like Biosciences (Churchill, 1989), Ecology (Li et al., 2001; Mari & Le Ber, 2006; Le Ber et al., 2006), Image (Pieczynski, 2003; Forbes & Pieczynski, 2009) and Signal processing (Rabiner & Juang, 1995). These stochastic models assume that the signals under investigation have a local property –called the Markov property– which states that the signal evolution at a given instant or around a given location is uniquely determined by its neighbouring values. In 1988, Pearl (Pearl, 1988) shown that these models can be viewed as specific dynamic Bayesian models which belong to a more general class called graphical models (Whittaker, 1990; Charniak, 1991).

The graphical models (GM) are the results of the marriage between the theory of probabilities and the theory of graphs. They represent the phenomena under study within graphs where the nodes are some variables that take their values in a discrete or continuous domain. Conditional –or causal– dependencies between the variables are graphically expressed. As an example, the relation between the random variables U, V and W depicted by Fig. 1 expresses that V and W are the reasons –more or less probable– of U. In a Bayesian attitude, the uncertainty about this relation is measured by the conditional probability $P(U/V,W)$ of observing U given V and W.

In graphical models, (see Fig. 2-4), some nodes model the phenomenon's data thanks to adequate distributions of the observations. They are called "observable" variables whereas the others are called "hidden" variables. The observable nodes of the graph give a frozen view of the phenomenon. In the time domain, the temporal changes are modelled by the set of transitions between the nodes. In the space domain, the theory of graphs allows to take into account the neighbourhood relations between the phenomenon's constituents.

The mining of temporal and / or spatial signals by graphical models can have several purposes:

Segmentation : in this task, the GM clusters the signal into stationary (or homogeneous) and transient segments or areas (Jain et al., 1999). The term stationnary means that the signal values are considered as independent outcomes of probability density functions (pdf). These areas are then post-processed to extract some valuable knowledge from the data.

Pattern matching : in this task, the GM measures the *a posteriori* probability $P(model = someLabel/observedData)$. When there are as many GM as labels, the best probability allows the classification of an unknown pattern by the label associated with the highest probability.

Background modelling : in order to make proper use of quantitative data, the GM is used as a background model to simulate an averaged process behavior that corrects for chance variation in the frequency counts (Huang et al., 2004). The domain expert compares the simulated and real data frequencies in order to distinguish if he / she is facing to over- or under-represented data that must be investigated more carefully.

In this chapter, we will present a general methodology to mine different kinds of temporal and spatial signals having contrasting properties: continuous or discrete with few or many modalities.
This methodology is based on a high order Markov modelling as implemented in a free software: CARROTAGE (see section 3). Section 2 gives the theoretical basis of the modelling. Section 3 describes a general flowchart for mining temporal and spatial signals using CARROTAGE. The next section is devoted to the description of three data mining applications following the same flowchart. Finally, we draw some conclusions in section 5.

2. The HMM as a graphical model

The Hidden Markov Model is a graphical model which represents the sequence of observations as a doubly stochastic process: an underlying "hidden" process, called the state sequence of random variables $Q_1, Q_2, ...Q_T$ and an output (observation) process, represented by the sequence $O_1, O_2, ...O_T$ over the same time interval (see Fig. 2-3). The sequence (Q_t) is a Markov chain and represents the different clusters that must be extracted.

2.1 HMM definition

We define a hidden Markov model by giving:

- $\mathbf{S} = \{s_1, s_2, \ldots, s_N\}$, a finite set of N states ;
- **A** a matrix defining the transition probabilities between the states:

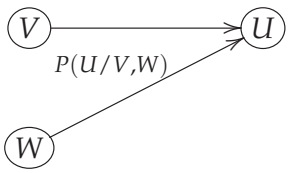


Fig. 1. Conditional dependency of U with V and W in a Bayesian network. The probability measures the confidence of the dependency

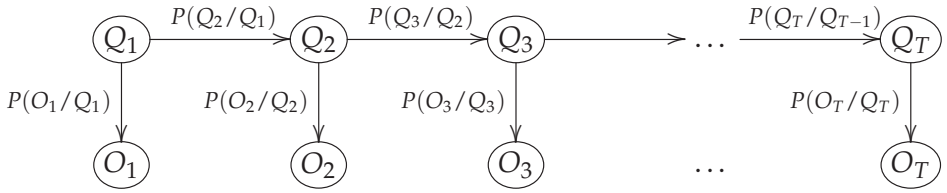


Fig. 2. Conditional dependencies in a HMM1 represented as a Bayesian network. The hidden variables (Q_t) govern the observable variables (O_t)

$\mathbf{A} = \left(a_{ij}\right)$ for a first order HMM (HMM1) (Fig. 2),

$\mathbf{A} = \left(a_{ijk}\right)$ for a second order HMM (HMM2) (Fig. 3);

– $\mathbf{b_i}(.)$ the distributions of observations associated to the states s_i. This distribution may be parametric, non parametric or even given by an HMM in the case of hierarchical HMM (Fine et al., 1998).

As opposite to a Markov chain where the states are unambiguously observed, in a HMM, the observations are not uniquely associated to a state s_i but are drawn from a random variable that has a conditional density $\mathbf{b_i}(.)$ that depends on the actual state s_i (Baker, 1974). There is a doubly stochastic process:

– the former is hidden from the observer, is defined on a set of states and is a Markov chain;

– the latter is visible. It produces an observation at each time slot –or index in the sequence– depending on the probability density function that is defined on the state in which the Markov chain stays at time t. It is often said that the Markov chain governs the latter.

2.2 Modelling the dependencies in the observable process

Defining the observation symbols is the first step of a HMM data processing. In this chapter, we will present our data mining work based on various GM applied on different kinds of signals having contrasting properties:

– genomic data characterized by long sequences (several millions) of the 4 nucleotides A, C, G, T (application 1);

– short temporal discrete sequences (around 10 value long) with a great number (around 50) of modalities like the temporal land use successions (LUS) of agricultural fields whose mosaic defines a 2-D spatial territory (application 2);

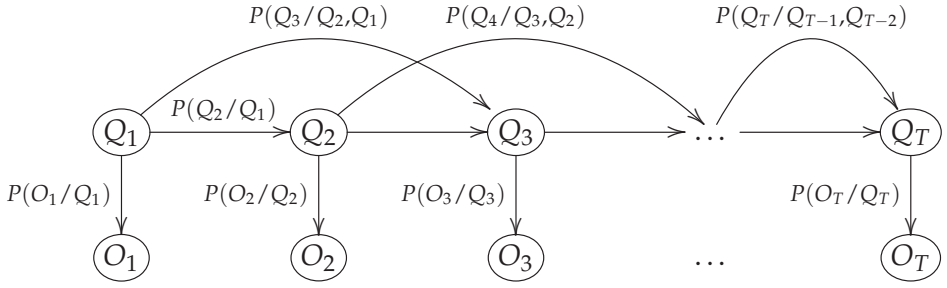


Fig. 3. Conditional dependencies in a HMM2 represented as a Bayesian network. The hidden variables (Q_t) govern the observable variables (O_t)

– continuous data like the values of a river width sampled from the river's source up to its end (application 3).

To take into account the correlations between successive or neighbouring observations, several options are possible.

2.2.1 Continuous observations

The usual way to model continuous random observations is to consider them as Gaussian distributed. When the observations are vectors belonging to $\mathbb{R}^d$, multivariate Gaussian `pdf` are used. The main reason of this consideration is that an unknown `pdf` can be approximated by a mixture of multivariate Gaussian `pdf`. To take into account the correlations between successive observations, first and second order regression coefficients (Furui, 1986) are stacked over the observation vector:

$$R(t) = \frac{\sum_{n=-n_0}^{n_0} nO(t+n)}{\sum_{n=-n_0}^{n_0} n^2} \tag{1}$$

where $O(t+n)$ is the observation (frame) $t+n$. The $2n_0+1$ frames involved in the computation of the regression coefficient $R(t)$ are centered around frame t. By this way, the vector at time t models the shape of the observation variations and incorporates information about the surrounding context.

2.2.2 Categorical observations

When the observations are discrete and belong to a finite set $C = \{c_1, c_2, ...c_M\}$, it is convenient to represent this correlation by adding new dependencies between the current observation and the previous observations. In the particular case shown in Fig. 4, the observation distribution is a conditional `pdf` $\mathbf{b_{iuv}}(o_t)$ that represents the conditional probability of observing o_t assuming the state s_i and the observations u and v that occured respectively at indices $t-1$ and $t-2$:

$$o_{t-1} = u,\ o_{t-2} = v \quad u, v \in C.$$

In the temporal domain, this leads to the definition of a `Mp-Mq HMM` where p is the order of the hidden Markov process and q refers to the dependencies in the observable process.
Another way to take into account the correlations between successive (neighbouring) observations, is to consider composite observations drawn from the n-fold product $C^n = C \times C \ldots C$. The elementary observation (for example, a nucleotide, a land use ...) is considered together with its context. This leads to the definition of *k-mer* (see section 4.1.1) in biology or land use succession in agronomy (see section 4.2.1.3). As a direct consequence, the `pdf` size will be changed from $|C|$ to $|C|^n$ where $|C|$ denotes the cardinality of C. It is

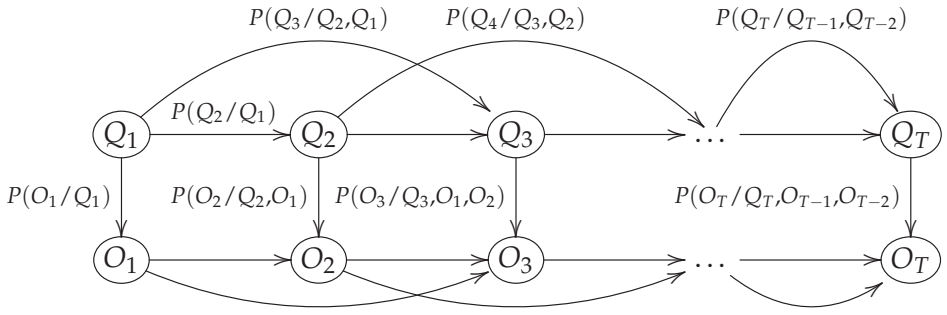


Fig. 4. Conditional dependencies of a `M2-M2 HMM` represented in a Bayesian network

then possible to control the balance between the parameter number assigned to the hidden variables and to the observable ones in the model.

2.3 Automatic estimation of a HMM2

The estimation of an HMM1 is usually done by the forward backward algorithm which is related to the EM algorithm (Dempster et al., 1977). We have shown in (Mari et al., 1997) that an HMM2 can be estimated following the same way. The estimation is an iterative process starting with an initial model and a corpus of sequences of observations that the HMM2 must fit even when the insertions, deletions and substitutions of observations occur in the sequences. The very success of the HMM is based on their robustness: even when the considered data do not suit a given HMM, its use can give interesting results. The initial model has equi-probable transition probabilities and a uniform distribution in each state. At each step, the forward backward algorithm determines a new model in which the likelihood of the sequences of observation increases. Hence this estimation process converges to a local maximum. Interested readers may refer to (Dempster et al., 1977; Mari & Schott, 2001) to find more specific details of the implementation of this algorithm.
If N is the number of states and T the sequence length, the second-order forward backward algorithm has a $N^3 \times T$ complexity for an HMM2.
The choice of the initial model has an influence on the final model obtained by convergence. To assess this last model, we use the Kullback-Leibler distance between the distributions associated to the states (Tou & Gonzales, 1974). Two states that are too close are merged and the resulting model is re-trained. Domain experts do not interfere in the process of designing a specific model, but they have a central role in the interpretation of the results that the final model gives on the data.

3. CARROTAGE a general framework to mine sequences

We have developed a knowledge discovery system based on high-order hidden Markov models for analyzing temporal data bases (Fig. 5). This system, named CARROTAGE[1], takes as input an array of discrete or continuous data –the rows represent the individuals and the columns the time slots– and builds a partition together with its *a posteriori* probability. CARROTAGE is a free software [2] under a Gnu Public License. It is written in C++ and runs under Unix systems. In all applications, the data mining processing based on CARROTAGE is decomposed into four main steps:

Model specification. Even if CARROTAGE may use models of any topology, we mainly use two different graph topologies: linear and ergodic. In a linear model, there is no circuit between the nodes except self loops on some nodes. Whereas in an ergodic model, all the nodes are inter connected; a node can reach all the others. The first HMM2 that CARROTAGE has to estimate is linear with equi-probable transitions from each state and uniform distributions of observations in every states. The only parameter let to the user is the number of states.

[1] CARROTAGE is a retro acronym that comes from the word carrot that can be translated by Markov in Russian and age to refer to the temporal component of the data. It is also a technique which consists in drilling a hole in some material (a tree or the ice of the Antarctic) to withdraw a cylinder that allows to date the process of creation

[2] http://www.loria.fr/~jfmari/App/

Iterative estimate of the model parameters. The parameter estimation of the model is performed by the forward backward algorithm for `M2-Md HMM`. Basically, given a sequence of symbols $\left(o_1^T\right) = o_1, o_2, \ldots o_T$ the second-order forward backward algorithm computes the expected count of the state transition $s_{i_1} \rightarrow s_{i_2} \rightarrow s_{i_3}$

$$\eta_t(i_1, i_2, i_3) = P(Q_{t-2} = s_{i_1}, Q_{t-1} = s_{i_2}, Q_t = s_{i_3} / O_1^T = o_1^T) \tag{2}$$

at index $t-2, t-1, t$.

The first parameter estimate is performed on a linear model to acquire a segmentation of the sequence into as many homogeneous regions than there are states in the specified model.

Linear to ergodic model transform. The estimated linear model is transformed into an ergodic one by keeping the previously estimated `pdf` and interconnecting the states. This allows the stochastic process to re-visit the states and, therefore, segment the data into an unconstrained number of homogeneous regions, each of them associated to a state.

Decoding. The decoding state uses the last iteration of EM algorithm to calculate the *a posteriori* probability of the hidden states. It is possible to compute three types of *a posteriori* probability. In all the following definitions, we assume that the hidden state s_i is attained at time t and that we have a T length observation sequence $\left(o_1^T\right)$.

type 0

$$P_0(i, t) = \sum_{i_1, i_2} \eta_t(i_1, i_2, i) \tag{3}$$

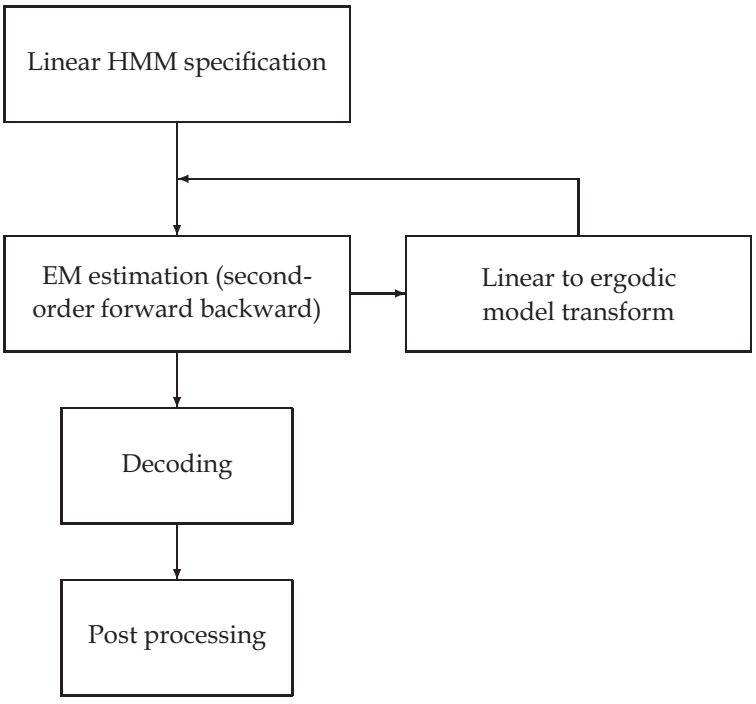


Fig. 5. General flow Chart of the data mining process using CARROTAGE

The *a posteriori* probability of the state s_i at index t assuming the whole sequence $\left(o_1^T\right)$.

type 1

$$P_1(i,t) = \sum_{i_1} \eta_t(i_1,i,i) \tag{4}$$

The *a posteriori* probability of the 2 state transition $s_i \to s_i$ at index t assuming the whole sequence $\left(o_1^T\right)$. This probability can be computed either by a `HMM1` or by a `HMM2`.

type 2

$$P_2(i,t) = \eta_t(i,i,i) \tag{5}$$

The *a posteriori* probability of the 3 state transition $s_i \to s_i \to s_i$ at index t assuming the whole sequence $\left(o_1^T\right)$. This probability is typical of a `HMM2`.

In some applications, as the mining of crop successions (see section 4.2), the *a posteriori* transition probability (type 1) between 2 states can be used and gives an interesting information. In such a case, we use:

$$P_1(i,j,t) = \sum_{i_1} \eta_t(i_1,i,j) \tag{6}$$

Post processing: The post processing is application dependent and involves mostly a classification step of the different segments. Further ad-hoc treatments must be performed in order to extract valuable information as shown in the application section.

4. Applications

4.1 Mining genomic data

In this section, we describe a new data mining method based on second-order HMM and combinatorial methods for Sigma Factor Binding Site (SFBS) prediction (Eng et al., 2009) and Horizontal Gene Transfer (HGT) (Eng et al., 2011) detection that voluntarily implements a minimum amount of knowledge. The original features of the presented methodology include (i) the use of the CARROTAGE framework, (ii) an automatic area extraction algorithm that captures atypical DNA motifs of various size based on the variation of the state *a posteriori* probability, and (iii) a set of post processing algorithms suitable to the biologic interpretation of these segments. On some points, our data mining method is similar to the work of Bize et al. (Bize et al., 1999) and Nicolas et al. (Nicolas et al., 2002). All the methods use one HMM to model the entire genome. The parameter estimation is done in all cases by the EM algorithm. All the methods look for attributing biological characteristics to the states by analyzing the state output *a posteriori* probability. But our method differs on the following points: we use (i) an HMM2 that has proved interesting capabilities in modelling short sequences, and (ii) depending on the modelled dependencies in the genomic sequence, we can locate either short nucleotides sequences that could be part of SFBS (box1 or box2) or more generally regulation sites for gene expression –Transcriptional Factor Binding sites (TFBS)– or even wider areas potentially acquired by HGT. These sequences are post processed to assess the exact nature of the heterogeneities (SFBS, TFBS or HGT).

4.1.1 Data preparation

In this application, the genome is modelled as an ordered nucleotide sequence whose unknown structure is represented by the state Markov chain. The index t in equation (2) refers to the nucleotide index in the ordered sequence of nucleotides. In a genome sequence, two templates must be considered depending upon the strength of the compositional biases. To incorporate the biased base composition of DNA strands relative to the position of the replication origin when a marked GC skew[3] is observed, as in the case of *Streptococcus thermophilus*, a sequence is constructed *in silico* by concatenating the two leading strands from the origin to the terminus of replication. Its reverse complement is also considered. In contrast, when the genome does not show a marked GC skew, as in *Streptomyces coelicolor*, the 5′ to 3′ sequence of the linear chromosome and its reverse complement are considered. In both cases, these two sequences are used for training purposes and specify two HMM2 named HMM2+ and HMM2-. The best decoding state is identified for both models.
We have also investigated the use of *k-mer* (Delcher et al., 1999) as output symbols instead of nucleotides. A *k-mer* may be viewed as a single nucleotide y_t observed at index t with a specific context $y_{t-k+1},\ldots y_{t-1}$ made of $k-1$ nucleotides that have been observed at index $t-k+1,\ldots,t-1$. Similarly, a DNA sequence can be viewed as a sequence of overlapping *k-mer* that an HMM analyzes with a consecutive shift of one nucleotide. For example, the seven nucleotide sequence `TAGGCTA` can be viewed as a sequence of seven *3-mer*: `##T` - `#TA` - `TAG` - `AGG` - `GGC` - `GCT` - `CTA`, where `#` represents an empty context.

4.1.2 *a posteriori* decoding

The mining of irregularities follows the general flow chart given in figure 5. The *a posteriori* probability variations look very different depending on the dependencies that are implemented in the genomic sequence. When modelling the *k-mer* sequence using a `M2-M0 HMM`, the decoding stage locates atypical short DNA segments (see Fig. 6) whereas the modelling of the nucleotide sequence using a `M2-M2 HMM` exhibits wider atypical areas (see Fig.7).

4.1.3 Post processing

The atypical regions extracted by the stochastic models must be processed in order to extract valuable information. A specific suite of algorithms has been designed and tuned in the two applications: TFBS and HGT detections.

4.1.3.1 TFBS retrieval

Our bacterial model is the Gram-positive actinomycete *Streptomyces coelicolor* whose genome is 8.7 Mb long. The streptomycetes are filamentous bacteria that undergo complex morphological and biochemical differentiation, both processes being inextricably interlinked. The purpose of the TFBS application is to retrieve composite motifs *box1-spacer-box2* involved in the *Streptomyces coelicolor* regulation. The two boxes can be part of the intergenic peak motifs (see Fig. 6). The spacer ranges from 3 to 25 and is tuned depending on the type of the investigated TFBS. The basic idea of the mining strategy is to cluster the set of intergenic ipeak motifs located by a `M2-M0 HMM` modelling *3-mer*, select a cluster having a well defined consensus, extend all the sequences belonging to this cluster and look for over-represented motifs by appropriate software (Hoebeke & Schbath, 2006). The consensus of the cluster acts

[3] the GC skew is a quantitative feature that measures the relative nucleotide proportion of G *versus* C in the DNA strand

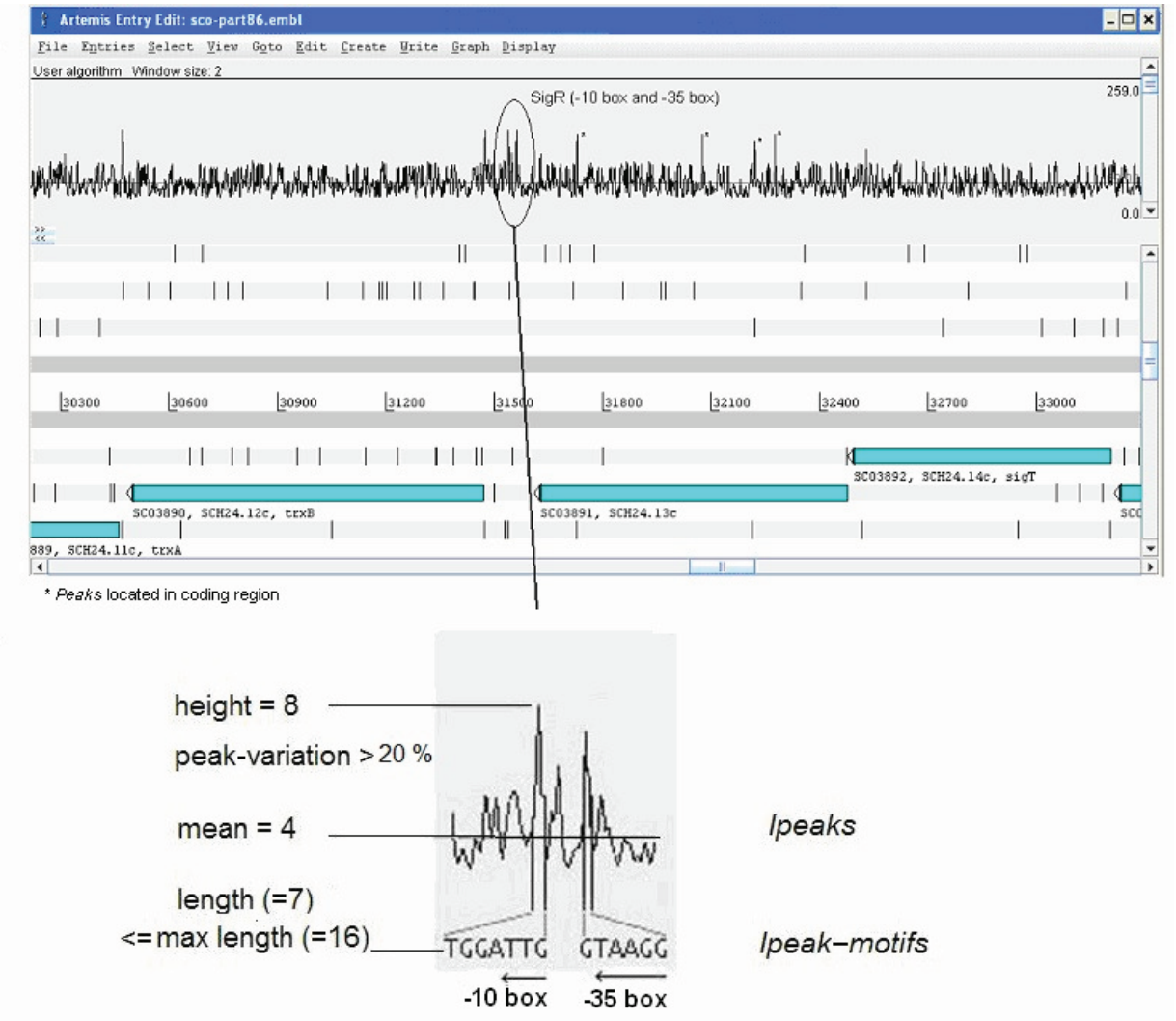


Fig. 6. *A posteriori* probability variation of a M2−M0 HMM hidden state as a function of the *3-mer* index in the *Streptomyces coelicolor* genome. The top graph shows the *a posteriori* probability together with the annotated physical sequence (using the EMBL file). As an example, among the intergenic peak motifs, the -35 box (GGAAT) and -10 box (GTT) motifs recognized by the sigma factor SigR are detected. Peak characteristics (peak-variation and length) are marked in the figure. The biological interpretation of the peaks inside the coding regions is not yet fully established (Eng et al., 2009)

for *box1*, the shorter motifs spaced with appropriate spacer value(s) act for *box2*. Interested readers will find in (Eng et al., 2009) an extensive description of this data mining strategy based on stochastic and combinatorial methods.

4.1.3.2 Horizontal gene transfer detection

Our bacterial model is the Gram-positive bacteria *Streptococcus thermophilus* which is a lactic acid bacteria carrying a 1.8 Mb genome and having a considerable economic importance. It is used as starter for the manufacturing of yogurts and cheeses. *Streptococcus thermophilus* is assumed to have derived very recently at the evolutionary time-scale (3,000-30,000 years back: the beginning of the pastoral epoch) from a commensal ancestor which is closely related to

Fig. 7. *A posteriori* probability variation of a M2−M2 HMM hidden state as a function of the nucleotide index in the *Streptococcus thermophilus* genome. The additional dependencies in the nucleotide sequence dramatically smooth the state *a posteriori* probability

the contemporary oral bacterium *Streptococcus salivarius* to adapt to its only known ecological niche: the milk. HGT deeply shaped the genome and played a major role in adaptation to its new ecological niche.

In this application, we have observed that the M2−M2 HMM modelling nucleotides performs better than M1−M2 HMM as implemented in SHOW software [4] (Nicolas et al., 2002) and M2−M0 HMM modelling *3-mer* (see section 4.1.3.1).

After tuning the HMM topology, the decoding state that captures the highest heterogeneities is selected by considering the distances between all states according to the Kullback-Leibler distance. The state which is the most far away from the others is selected. On this state, the variations of the *a posteriori* probability as a function of the index in the nucleotide sequence are analyzed. The positions having *a posteriori* probabilities higher than the mean over the whole genome are considered. Regions enriched in these positions through at least 1000 nucleotide length were extracted and named atypical regions. A total of 146 atypical regions were extracted. If a gene were at least half included in these regions then it was considered. A total of 362 genes of 1915 (the whole gene set of the bacterium), called "atypical", were retrieved from these regions. Based on their functional annotation and their sporadic distribution either at the interspecific level (among the other genomes belonging to the same phylum: the Firmicutes) or at the intraspecific level (among a collection of 47 strains of *Streptococcus thermophilus*), a HGT origin can be predicted for a large proportion (about two thirds) (Eng, 2010).

4.2 Mining agricultural landscapes

In agricultural landscapes, land-use (LU) categories are heterogeneously distributed among different agricultural fields managed by farmers. At a first glance, the landscape spatial organization and its temporal evolution seem both random. Nevertheless, they reveal the presence of logical processes and driving forces related to the soil, climate, cropping system, and economical pressure. The mosaic of fields together with their land-use can be seen as a noisy picture generated by these different processes.

Recent studies (Le Ber et al., 2006; Castellazzi et al., 2008) have shown that the ordered sequences of LU in each field can be adequately modelled by a high order Markov process. The LU at time t depends upon the former LU at previous times: $t-1, t-2 \ldots$ depending on

[4] http://genome.jouy.inra.fr/ssb/SHOW/

	Case study	
	Niort Plain	**Yar watershed**
Data source	Land-use surveys	Remote sensing
Surface (sq. km)	350	60
Study period	1996 to 2007	1997 to 2008
Number of LU modalities	47	6
Spatial representation	Vector	Raster (converted to vector)
Elementary spatial entities	Elementary plots (polygons)	Pixels (20 x 20 sq. m)
Data base format	ESRI Shapefile	ESRI Shapefile

Table 1. Comparison between 2 land-use databases coming from two different sources: land-use surveys and remote sensing

the order of the Markov process. In the space domain, the theory of the random Markov fields is an elegant mathematical way for accounting neighbouring dependencies (Geman & Geman, 1984; Julian, 1986). In this section, we present a data mining method based on CARROTAGE to cluster a landscape into patches based on its pluri annual LU organization. Two medium-size agricultural landscapes will be considered coming from different sources: long-term LU surveys or remotely sensed LU data.

4.2.1 Data preparation

For CARROTAGE, the input corpus of LU data is an array in which the columns represent the LU year by year and the rows represent regularly spaced locations in the studied landscape (e.g. 1 point every 20 m). Data preparation aims at reducing the requirement of the memory resources while putting the data in the appropriate format required by CARROTAGE. The data preparation process must tackle several issues: (i) to regroup into LU categories the different LU when there are too many observations, (ii) to define the elementary observation for the HMM, and (iii) to choose the sampling spatial resolution.

The corpus of spatiotemporal LU data is generally built either from long-term LU surveys or from remotely sensed LU data. Depending on the data source, several differences in the LU database may exist. These differences are mostly regarding the number of LU modalities and the representation of the spatial entities: polygons in vector data or pixels in raster data. In the following, the first data source (long-term LU field surveys) is illustrated by the Niort Plain case study (Lazrak et al., 2010), and the second (remotely sensed LU) is illustrated by the Yar watershed case study. Principal characteristics of the two case studies are summarized in table 1.

4.2.1.1 The agricultural landscape mosaic

The agricultural landscape can be seen as an assemblage of polygons of variable size where each polygon holds a given LU. When data derives from LU surveys, the polygons are fields bounded by a road, a path or a limit of a neighbouring field. The polygon boundaries can change every year. To take into account this change, the surveyors update each year the boundaries of fields in the GIS database. For remotely sensed images, the polygons are obtained by grouping similar pixels in the same class and are represented in vector format. In the two cases, the list of the polygon boundaries –that change over the time– led to the definition of the elementary polygon –the plot– as the result of the spatial union of previous polygon boundaries (Figure 8). Each plot holds one LU succession during the study period. There are about 20,000 elementary plots in the Niort study area over the 1996 – 2007 period.

The corpus of land-use data is next sampled and is represented in a matrix in which the columns are related to the time slots and the rows to the different grid locations.
Following Benmiloud and Pieczynski (Pieczynski, 2003), we have approximated the Markov random field (MRF) by sampling the 2-D landscape representation using a regular grid and, next, defining a scan by a Hilbert-Peano curve (figure 9). The Markov field is then represented by a Markov chain. Two successive points in the Markov chain represent two neighbour points in the landscape but the opposite is not true, nevertheless, this rough modelling of the neighbourhood dependencies has shown interesting results compared to an exact Markov random field modelling (Benmiloud & Pieczynski, 1995). To take into account the irregular neighbour system, we can also adjust the fractal depth to the mean plot size. The figure 9 illustrates this concept.

4.2.1.2 LU categories definition

When LU derive from LU surveys, there is often a great number of LU modalities which must be reduced by defining LU categories. For the Niort Plain case study, the 47 LU have been grouped with the help of agricultural experts in 10 categories (see Tab. 2) following an

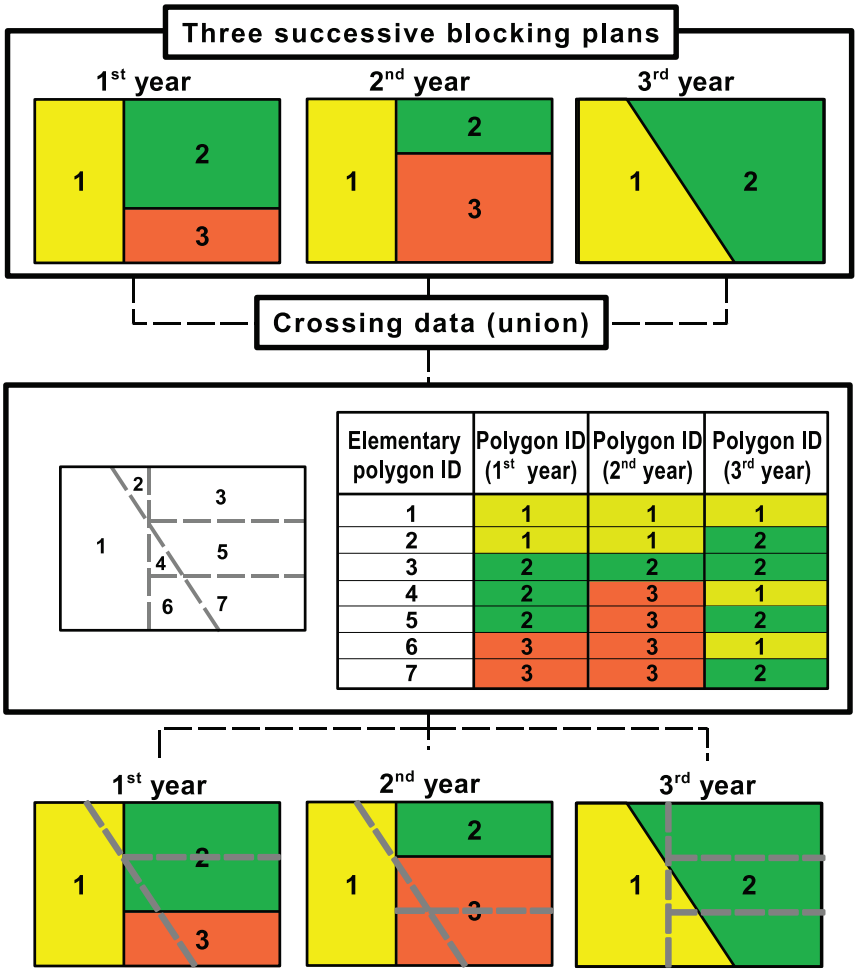


Fig. 8. An example of field boundary evolution over three successive years. The union of field boundaries during this period leads to the definition of seven plots

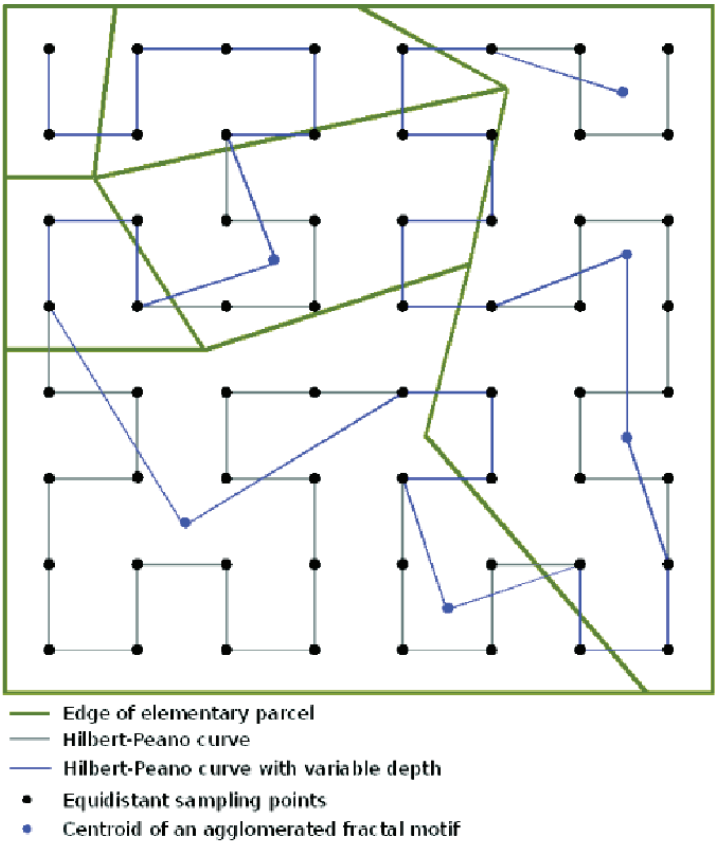


Fig. 9. Variable depth Hilbert-Peano scan to take into account the field size. Two successive merging in the bottom left field yield to the agglomeration of 16 points

approach based on the LU frequency in the spatiotemporal database and the similarity of crop management.
For the Yar watershed case study, only six LU have been distinguished: Urban, Water, Forest, Grassland, Cereal and Maize. There was no need of grouping them into categories.

4.2.1.3 Choice of the elementary observation

An elementary observation can range from a LU (such as Cereal in the Yar watershed case study) or a LU category (such as Wheat in the Niort Plain case study) to a LU succession (LUS) spanning several years. For this latter, the length of the LU succession influences the interpretation of the final model. However, the total number of LUS is a power function of the succession length, and memory resources required during the estimation of HMM2 parameters increase dramatically.
To determine the succession length, we compared the diversity of LUS between field-collected data (the Niort Plain) and randomly generated data for different lengths of successions (Fig. 10(a)). For this case study, 4-year successions begin to clearly differentiate the landscape from a random landscape in which the LU are randomly allocated in the plots. Therefore, 4-year successions appear to be the shortest HMM2 elementary observation symbol suitable for modelling LUS within the Niort Plain landscape. The choice for the elementary observation can also be set by domain specialists based on previous works (Le Ber et al., 2006; Mignolet et al., 2007). This was the case for the Yar watershed where we chose to model the

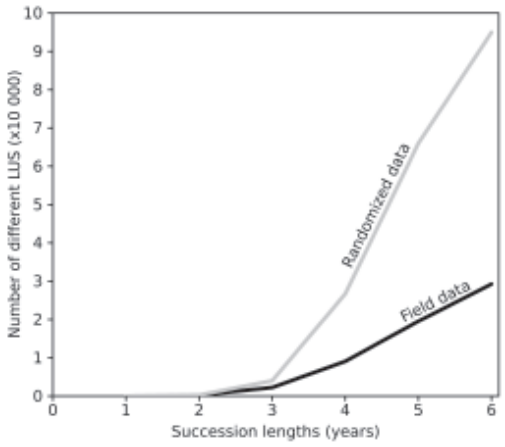
LU category	LU	Frequency	Cumul
Wheat	Wheat, bearded wheat, cereal	0.337	0.337
Sunflower	Sunflower, ryegrass followed by sunflower	0.139	0.476
Rapeseed	Rapeseed	0.124	0.600
Urban	Built area, peri-village, road	0.096	0.696
Grassland	Grassland of various types, alfalfa,...	0.078	0.774
Maize	Maize, ryegrass followed by maize	0.076	0.850
Forest	Forest or hedge, wasteland	0.034	0.884
Winter barley	Winter barley	0.034	0.918
Ryegrass	Ryegrass, ryegrass followed by ryegrass	0.024	0.942
Pea	Pea	0.022	0.964
Others	Spring barley, grape vine, clover, field bean, ryegrass, cereal-legume mixture, garden/market gardening,...	0.036	1.000

Table 2. Composition and average frequencies of adopted LU categories (Lazrak et al., 2010)

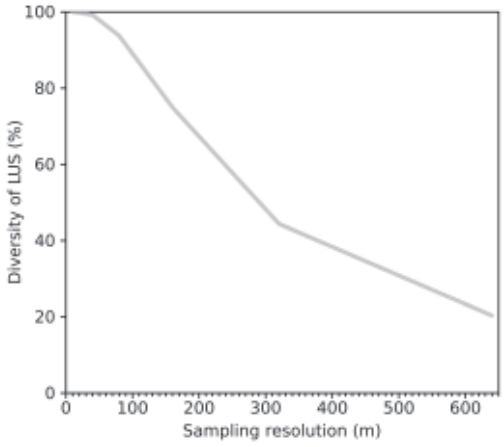
agricultural dynamics through 3-year LUS.

4.2.1.4 Choice of the spatial resolution

For medium-size and large landscapes, a high-resolution sampling generates a large amount of data. With such amount, only rough models can be tested. On the other hand, with a coarse resolution sampling, small fields are omitted. In order to have an objective criterion for choosing the optimal spatial resolution, we can estimate information loss in terms of LUS diversity for increasingly coarse resolution samplings. Figure 10(b) shows the obtained curve for the Niort Plain case study. The tested resolutions were: 10, 20, 40, 80, 160, 320 and 640 m. Irregularity in sampling intervals is dictated by an algorithmic constraint: the resolution must be proportional to a power of 2. The most precise resolution is considered as the reference



(a) Compared diversity of LUS between field-collected data and 10 random generated data sets for different succession lengths



(b) Information loss in terms of LUS diversity in relation to sampling resolutions for 4-year LUS

Fig. 10. Relations between LUS diversity and sampling rates

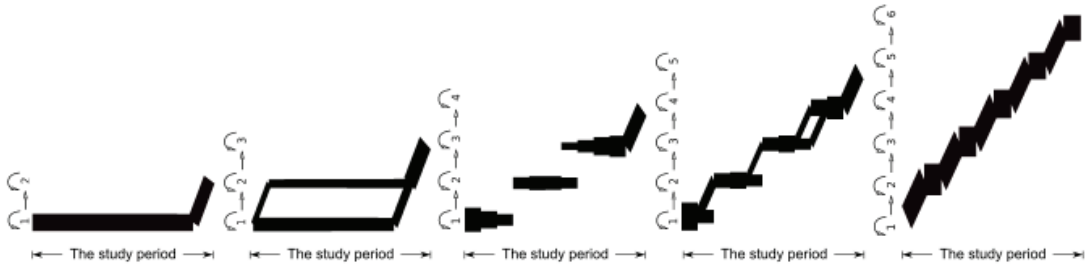


Fig. 11. Seeking the best temporal segmentation of the Yar watershed study period by using 5 growing state number linear HMM2. The line width is proportional to the *a posteriori* transition probability (Eq. 6). The 6 state HMM2 segments the study period into 6 non-overlapping periods

(100%). As a compromise, we chose the 80 m x 80 m resolution that led to a corpus 64 times smaller than the original one, with only a loss of 6% in information diversity.
For the Yar watershed landscape, which has a surface roughly 7 times smaller than the Niort Plain landscape and has few LU modalities, we were not constrained by the corpus size. Thus, we chose a 20 m x 20 m resolution which was the original resolution of satellite images used to identify the LU.

4.2.2 *a posteriori* decoding

We propose to build a time spatial analysis through spatial analysis of crop dynamics. This data mining method is a time x space analysis where a temporal analysis is performed in order to identify temporal regularities before locating these regularities in the landscape by means of a hierarchical HMM2 (HHMM2). The HHMM2 allows segmenting the landscape into patches, each of them being characterized by a temporal HMM2.

4.2.2.1 Mining temporal regularities

Depending on the investigated temporal regularities, we can either use a linear HMM2 or a multi-column ergodic HMM2 (Fig. 12). Linear models allow segmenting the study period into homogeneous sub-periods in terms of LUS distributions (see Figure 11).
Multi-column ergodic models (Mari & Le Ber, 2006; Le Ber et al., 2006) (Fig. 12) have been designed for measuring the probability of a succession of land-use categories. Actually, we have defined a specific state, called the *Dirac state*, whose distribution is zero except on a particular land-use category. Therefore, the transition probabilities between the Dirac states measure the probabilities between the land-use categories. Figure 12 shows the topology of a `HMM2` that has two kinds of states: Dirac states associated to the most frequent land-use categories (wheat, sunflower, barley, ...) and *container states* associated to uniform distributions over the set of observations. The estimation process usually empties the container state of the land-use categories associated with Dirac states. Therefore this model generalises both hidden Markov models and Markov models.
The model generation follows the same flowchart given in figure 5. When it is needed, the *Dirac states* can be initialized by some search patterns for capturing one or many particular observations.
Agronomists interpret the resulting diagrams to find the LU dynamics. Figure 13 shows a quasi steady agricultural system. The crop rotations involve Rapeseed, Sunflower and Wheat. In order to determine the exact rotations (2-year or 3-year), it is necessary to envisage the

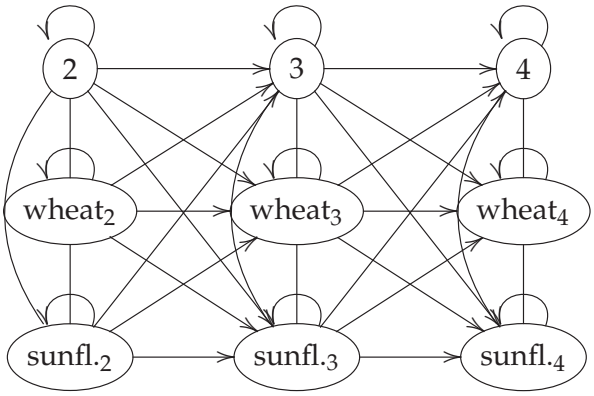


Fig. 12. Multiple column ergodic model: the states denoted 2, 3 and 4 are associated to a distribution of land-use categories, as opposite to the *Dirac* states denoted with a specific land-use category. The number of columns determines the number of time intervals (periods). A connection without arrow means a two directional connection

modelling of 4-year LUS (Lazrak et al., 2010). Note the monoculture of Wheat that starts in 2004.

4.2.2.2 Spatial clustering based on HMM2

We model the spatial structure of the landscape by a MRF whose sites are random LUS. The dynamics of these LUS are modelled by a temporal HMM2. This leads to the definition of

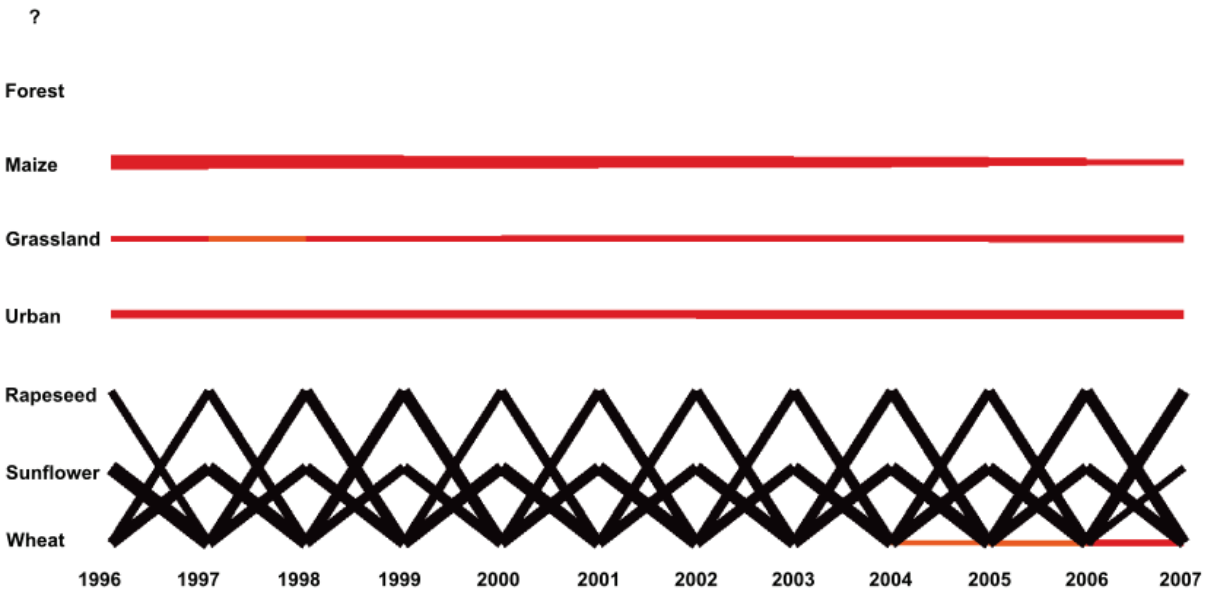


Fig. 13. Markov diagram showing transitions between LU categories in the Niort Plain. The x-axis represents the study period. The y-axis stands for the states of the ergodic one-column HMM2 used for data mining. Each state represents one LU category. The state '?' is the *container* state associated to a `pdf`. Diagonal transitions stand for inter-annual LU changes. Horizontal transitions indicate inter-annual stability. For simplicity, only transitions whose frequencies are greater than 5 % are displayed. The line width reflects the *a posteriori* probability of the transition assuming the observation of the 12-year LU categories (Eq. 6)

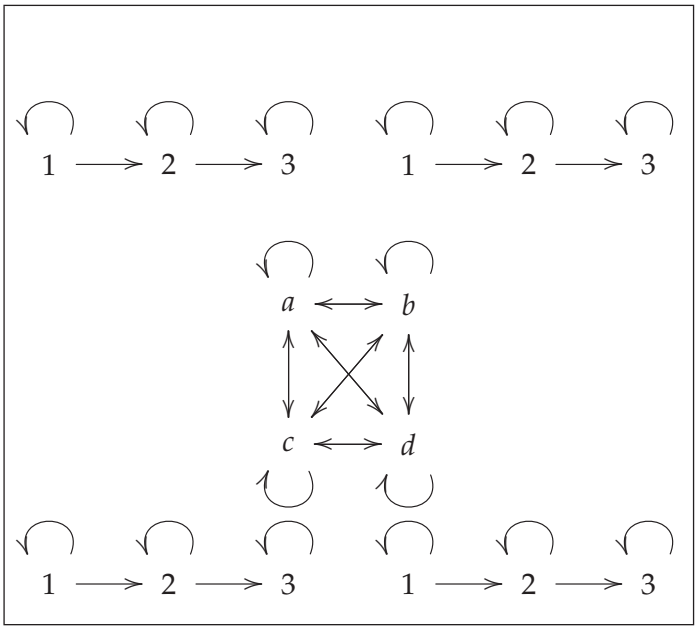


Fig. 14. Example of hierarchical HMM2. Each spatial state a, b, c, d of the master HHMM2 (ergodic model) is a temporal HMM2 (linear model) whose states are 1, 2, 3

a hierarchical HMM2 (Figure 14) where a master HMM2 approximates the MRF. Then, the probability of LUS is given by a temporal HMM2 as fully described in (Fine et al., 1998; Mari & Le Ber, 2006; Lazrak et al., 2010). This hierarchical HMM is used to segment the landscape into patches, each of them being characterized by a temporal HMM2. At each index l in the Hilbert-Peano curve, we look for the best *a posteriori* state in the HHMM2 (Maximum Posterior Mode algorithm). The state labels, together with the geographic coordinates of the indices l, determine a clustered image of the landscape that can be coded within an ESRI shapefile. An example of this segmentation for the Yar watershed case study is given in Figure 15.

4.2.3 Post processing

For the Yar watershed case study, we have performed preliminary temporal segmentation tests with linear models having an increasing number of states (Figure 11). This led us to use a 6-state HMM2 to segment the study period into 6 sub-periods characterized by different `pdf`. Plotting together the 6 sub-periods gives a global view on the LU dynamics (Figure 15).
In figure 15, the Yar watershed is represented by a mosaic of patches of LU evolutions. These patches are associated to a 5-state ergodic HHMM2. States 1 and 2, respectively represent Forest and Urban and are steady during the study period. The Urban state is also populated by less frequent LU that constitute its privileged neighbours. Grassland is the first neighbour of Urban, but it vanishes over the time. The other 3 states exhibit a greater LU diversity and a more pronounced temporal variation. In state 3, Grassland, Maize and Cereal evolve together until the middle of the study period. Next, Grassland and Maize decrease and are replaced by Cereal. This trend shows very likely that a change of cropping system was undertaken in the patches belonging to this state.

4.3 Mining hydro-morphological data

In this section we describe the use of `HMM2` for the segmentation of data describing river channels. Actually, a river channel is considered as a continuum and is characterised

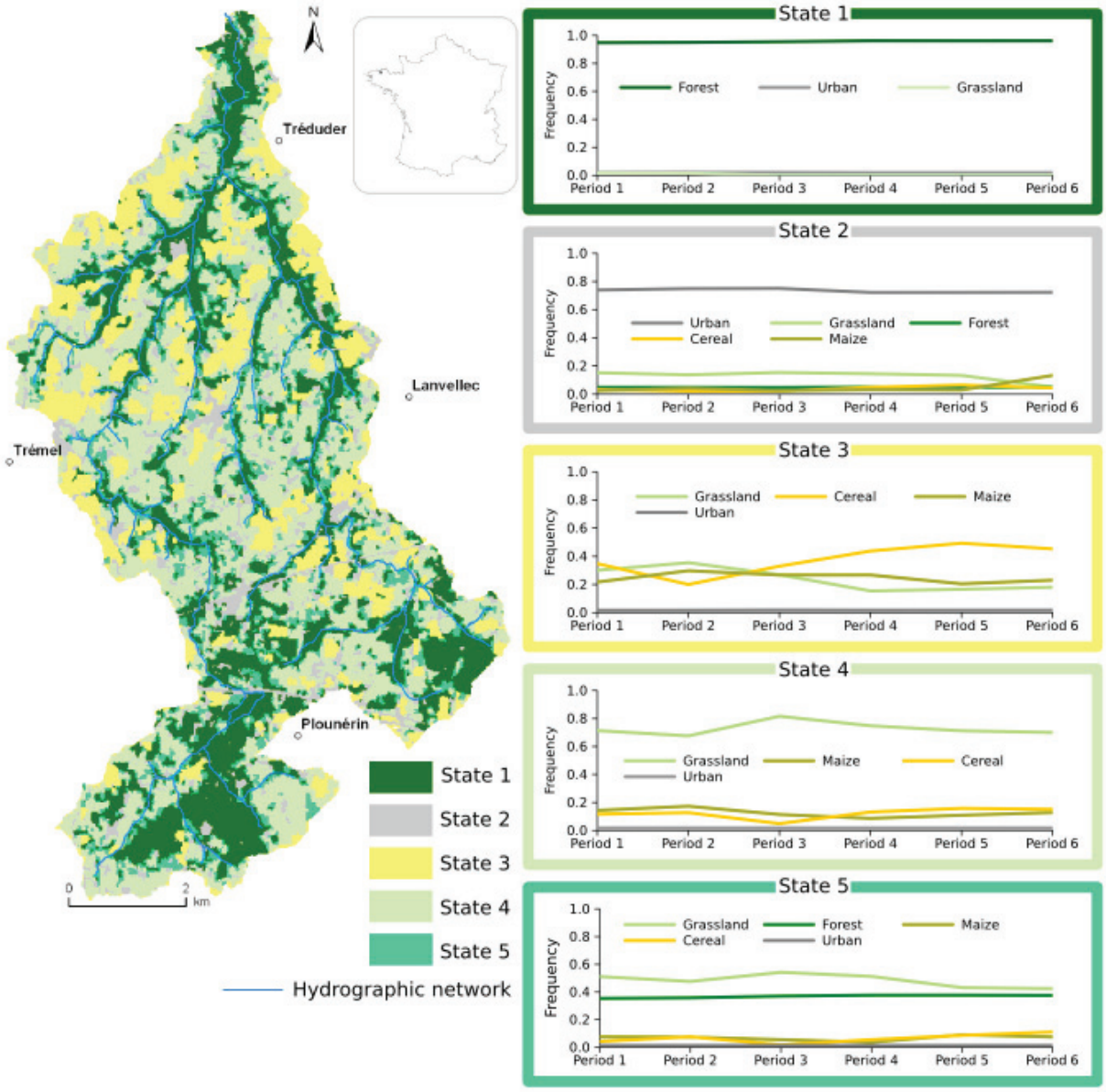


Fig. 15. The Yar watershed seen as patches of LU dynamics. Each map unit stands for a state of the HHMM2 used to achieve the spatial segmentation. Each state is described by a diagram of the LU evolution. The 6 sub-periods are the time slots derived from the temporal segmentation with the 6-state HMM2 describing each state of the HHMM2. Location of the Yar watershed in France is shown by a black spot depicted in the upper middle box

by its width or depth that is increasing downstream whereas its slope and grain size decrease (Schumm, 1977). The segmentation of this continuum with respect to local characteristics is an important issue in order to better manage the river channels (e.g. protection of plant or animal species, prevention of flood or erosion processes, etc.). Several methods have been proposed to perform such a segmentation. Markov chains Grant et al. (1990) and `HMM1` (Kehagias, 2004) are also been used.

4.3.1 Data preparation

The aim is to establish homogeneous units of the river Drome (South-East of France) continuum according to its geomorphological features. First of all, the continuum has been segmented within 406 segments of 250 meters length. Each segment is then described with several variables computed from aerial photographs (years 1980/83 and 1994/96) supplemented with terrain observations. Details about the computing of these variables can be found in (Aubry & Piégay, 2001; Alber & Piégay, 2010; Alber, 2010). In the following, we focus on the variable describing the width of the active channel (i.e. the water channel and shingle banks without vegetation).

4.3.2 *a posteriori* decoding

The stochastic modelling follows the same flow chart given in Fig. 5. Both linear and ergodic models have been used. The `pdf` associated in the `M2-M0 HMM` are univariate Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$.

$$b_i(O_t) = \mathcal{N}(O_t; \mu_i, \Sigma_i) \tag{7}$$

where O_t is the input vector (the frame) at index t and $\mathcal{N}(O_t; \mu, \Sigma)$ the expression of the likelihood of O_t using a gaussian density with mean μ and variance Σ. The maximum likelihood estimates the mean and covariance are given by the formulas using the definition of P_0 (cf. Equ.3):

$$\overline{\mu_i} = \frac{\sum_t P_0(i,t) O_t}{\sum_t P_0(i,t)} \tag{8}$$

$$\overline{\Sigma_i} = \frac{\sum_t P_0(i,t)(O_t - \mu_i)(O_t - \mu_i)^t}{\sum_t P_0(i,t)} \tag{9}$$

Specific user interfaces have been designed, in order to fit the experts' requirements: the original data are plotted, together with the mean value and the standard deviation of the current (most probable) state.
The linear model (Fig. 16) allows to detect a limited number (due to the specified number of states) of high variations, i.e. large and short vs narrow and long sections of the river channel. The ergodic model (Fig. 17) allows to detect an unknown number of small variations and repetitions.

4.3.3 Post processing

The final aim of this study is to build a geomorphical typology based on the river characteristics and to link it to external criteria (e.g. geology, land-use). The clustering is useful to define a relevant scale for this typology. If the typology is limited to the Drome river, the linear `HMM` allows to detect a set of segments that can be characterised by further variables and used as a basis for the typology. Ten segments for 101.5 kilometres appeared to be a good

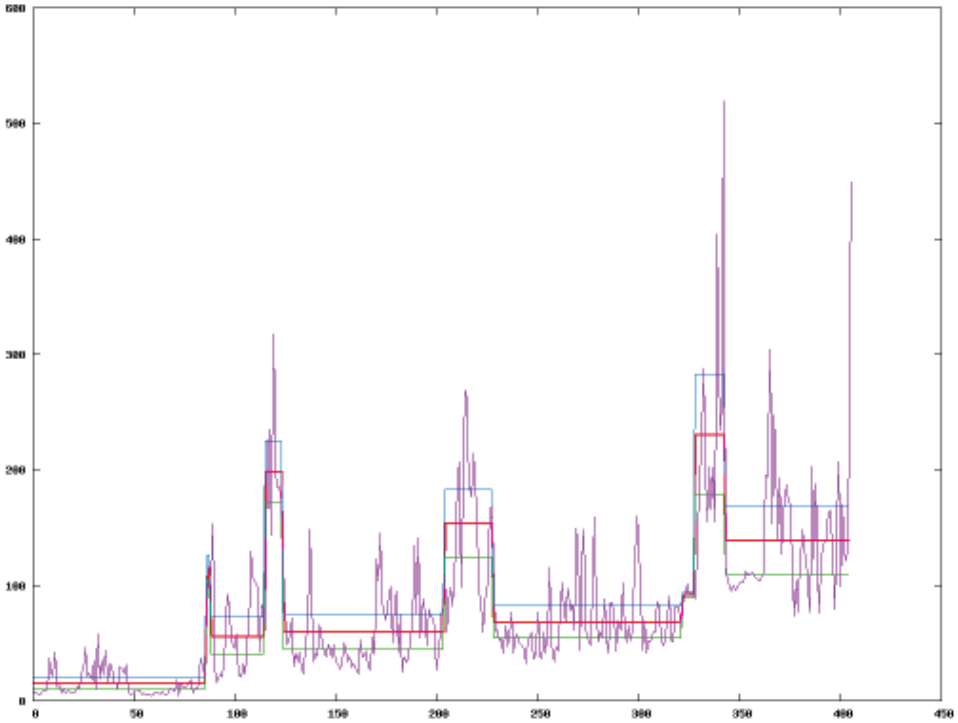


Fig. 16. Clustering the active channel width of the Drome river: linear HMM2 with 10 states

scale. On the contrary, if a whole network is considered -with several rivers and junctions-, the segmentation performed by the ergodic HMM would be more interesting since it allows to segment the data with less states than the linear model and to reveal similar zones (i.e.

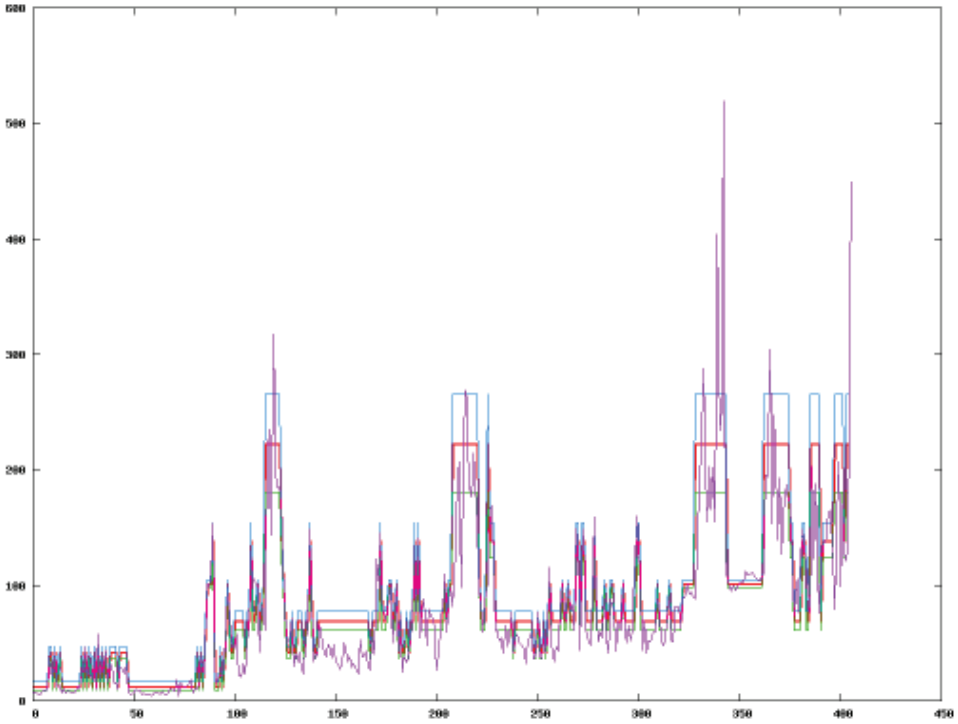


Fig. 17. Clustering the active channel width of the Drome river: ergodic HMM2 with 6 states

belonging to the same state) in the network. The probability transitions between states can also be exploited to reveal similar sequences of states along the network and thus to perform nested segmentations. Furthermore, transition areas appearing as significant mixtures of several states may be dealt with separately or excluded from a typology. Specific algorithms have to be designed and tuned to deal with these last questions.

5. Conclusions

We have described in this chapter a general methodology to mine temporal and spatial data based on a high order Markov modelling as implemented in CARROTAGE. The data mining is basically a clustering process that voluntary implements a minimum amount of knowledge. The HMM maps the observations into a set of states generated by a Markov chain. The classification is performed, both in time domain and spatial domain, by using the *a posteriori* probability that the stochastic process stays in a particular state, assuming a sequence of observations. We have shown that spatial data may be re-ordered using a fractal curve that preserves the neighbouring information. We adopt a Bayesian point of view and measure the temporal and the spatial variability with the *a posteriori* probability of the mapping. Doing so, we have a coherent processing both in temporal and spatial domain. This approach appeared to be valuable for time space data mining.
In the genomic application, two different HMM (M2−M0 HMM and M2−M2 HMM) have extracted meaningful regularities that are of interest in the area of promoter and HGT detection. The dependencies in the observation sequence smooth dramatically the *a posteriori* probability. We put forward the hypothesis that this smoothing effect is due to the additional normalisation constraints used to transform a 64 bin pdf of *3-mer* into 16 pdf of nucleotides. This smoothing effect allows the extraction of wider regularities in the genome as it has been shown in the HGT application.
In the agronomic application, the hierarchical HMM produces a time space clustering of agricultural landscapes based on the LU temporal evolution that gives to the agronomist a concise view of the current trends. CARROTAGE is an efficient tool for exploring large land use databases and for revealing the temporal and spatial organization of land use, based on crop sequences (Mari & Le Ber, 2003). Furthermore, this mining strategy can also be used to investigate and visualize the crop sequences of a few specific farms or of a small territory. In a recent work (Schaller et al., 2010) aiming at modelling the agricultural landscape organization at the farm and landscape levels, the stochastic regularities have been combined with farm surveys to validate and explain the individual farmer decision rules. Finally, the results of our analysis can be linked to models of nitrate flow and used for the evaluation of water pollution risks in a watershed (**?**).
In the mining of hydro-morphological data, the HMM have given promising results. They could be used to perform nested segmentations and reveal similar zones in the hydrological network. We are carrying out extensive comparisons with other methods in order to assess the gain given by the high order of the Markov chain modelling.
In all these applications, the extraction of regularities has been achieved following the same flowchart that starts by the estimation of a linear HMM to get initial seeds for the probabilities and, next, a linear to ergodic transform followed by a new estimation by the forward backward algorithm. Even if the data do not suit the model, the HMM can give interesting results allowing the domain specialist to put forward some new hypothesis. Also, we have noticed that the data preparation is a time consuming process that conditions all further steps

of the data mining process. Several ways of encoding elementary observations have been tried in all applications during our interactions with the domain specialists.
A much discussed problem is the automatic design of the HMM topology. So far, CARROTAGE does not implement any tools to achieve this goal. We plan to improve CARROTAGE by providing it with these tools and assess this new feature in the numerous case studies that we have already encountered. Another new trend in the area of artificial intelligence is the clustering of both numerical and symbolic data. Also, based on their transition probabilities and pdf, the HMM could be considered as objects that have to be compared and clustered by symbolical methods. The frequent items inside the pdf can be analyzed by frequent item set algorithms to achieve a description of the intent of the classes made of the most frequent observations that have been captured in each state in the HMM. These issues must be tackled if we want to deal with different levels of description for large datasets.

6. Acknowledgments

Many organizations had provided us with support and data. The genetic data mining work was supported by INRA, the région Lorraine and the ACI IMP-Bio initiative. Hydro-morphological data were provided by H. Piégay and A. Alber, UMR 5600 CNRS, Lyon. The original idea of this work arose from discussions with T. Leviandier, ENGEES, Strasbourg. The agronomic work was supported by the ANR-ADD-COPT project, the API-ECOGER project, the région Lorraine and the ANR-BiodivAgrim project. We thank the two CNRS teams: UPR CEBC (Chizé) for their data records obtained from the "Niort Plain database" and UMR COSTEL (Rennes) for the "Yar database".

7. References

Alber, A. (2010). PhD thesis, U. Lyon 2, France. to be published.

Alber, A. & Piégay, H. (2010). Disaggregation-aggregation procedure for characterizing spatial structures of fluvial networks: applications to the Rhône basin (France), *Geomorphology* . In press.

Aubry, P. & Piégay, H. (2001). Pratique de l'analyse de l'autocorrélation spatiale en géomorphologie fluviale : définitions opératoires et tests, *Géographie Physique et Quaternaire* 55(2): 115–133.

Baker, J. K. (1974). Stochastic Modeling for Automatic Speech Understanding, *in* D. Reddy (ed.), *Speech Recognition*, Academic Press, New York, New-York, pp. 521 – 542.

Benmiloud, B. & Pieczynski, W. (1995). Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images, *Traitement du signal* 12(5): 433 – 454.

Bize, L., Muri, F., Samson, F., Rodolphe, F., Ehrlich, S. D., Prum, B. & Bessières, P. (1999). Searching Gene Transfers on *Bacillus subtilis* Using Hidden Markov Models, *RECOMB'99*.

Castellazzi, M., Wood, G., Burgess, P., Morris, J., Conrad, K. & Perry, J. (2008). A systematic representation of crop rotations, *Agricultural Systems* 97: 26–33.

Charniak, E. (1991). Bayesian Network without Tears, *AI magazine* .

Churchill, G. (1989). Stochastic Models for Heterogeneous DNA Sequences, *Bull Math Biol* 51(1): 79 – 94.

Delcher, A., Kasif, S., Fleischann, R., Peterson, J., White, O. & Salzberg, S. (1999). Alignment of whole genomes, *Nucl. Acids Res.* 27(11): 2369 – 2376.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum-Likelihood From Incomplete Data Via

The EM Algorithm, *Journal of Royal Statistic Society, B (methodological)* 39: 1 – 38.

Eng, C. (2010). *Développement de méthodes de fouille de données fondées sur les modèles de Markov cachés du second ordre pour l'identification d'hétérogénéités dans les génomes bactériens*, PhD thesis, Université Henri Poincaré Nancy 1. http://www.loria.fr/~jfmari/ACI/these_eng.pdf.

Eng, C., Asthana, C., Aigle, B., Hergalant, S., Mari, J.-F. & Leblond, P. (2009). A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods, *Journal of Computational Biology* 16(9): 1211–1225. http://hal.inria.fr/inria-00419969/en/.

Eng, C., Thibessard, A., Danielsen, M., Rasmussen, T., Mari, J.-F. & Leblond, P. (2011). In silico prediction of horizontal gene transfer in *Streptococcus thermophilus*, *Archives of Microbiology* . in preparation.

Fine, S., Singer, Y. & Tishby, N. (1998). The Hierarchical Hidden Markov Model: Analysis and Applications, *Machine Learning* 32: 41 – 62.

Forbes, F. & Pieczynski, W. (2009). New Trends in Markov Models and Related Learning to Restore Data, *IEEE International Workshop on Machine Learning for Signal Processing (MSLP)*, IEEE, Grenoble.

Forney, G. (1973). The Viterbi Algorithm, *IEEE Transactions* 61: 268–278.

Furui, S. (1986). Speaker-independent Isolated Word recognition Using Dynamic Features of Speech Spectrum, *IEEE Transactions on Acoutics, Speech and Signal Processing* .

Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 6.

Grant, G., Swanson, F. & Wolman, M. (1990). Pattern and origin of stepped-bed morphology in high-gradient streams, Western Cascades, Oregon, *Geological Society of America Bulletin* 102: 340–352.

Hoebeke, M. & Schbath, S. (2006). R'mes: Finding exceptional motifs. user guide, *Technical report*, INRA.
URL: *http://genome.jouy.inra.fr/ssb/rmes*

Huang, H., Kao, M., Zhou, X., Liu, J. & Wong, W. (2004). Determination of local statistical significance of patterns in markov sequences with application to promoter element identification, *Journal of Computational Biology* 11(1).

Jain, A., Murty, M. & Flynn, P. (1999). Data Clustering: A Review, *ACM Computing Surveys* 31(3): 264 – 322.

Julian, B. (1986). On the Statistical Analysis of Dirty Picture, *Journal of the Royal Statistical Society* B(48): 259 – 302.

Kehagias, A. (2004). A hidden Markov model segmentation procedure for hydrological and environmental time series, *Stochastic Environmental Research* 18: 117–130.

Lazrak, E., Mari, J.-F. & Benoît, M. (2010). Landscape regularity modelling for environmental challenges in agriculture, *Landscape Ecology* 25(2): 169 – 183. http://hal.inria.fr/inria-00419952/en/.

Le Ber, F., Benoît, M., Schott, C., Mari, J.-F. & Mignolet, C. (2006). Studying Crop Sequences With CarrotAge, a HMM-Based Data Mining Software, *Ecological Modelling* 191(1): 170 – 185. http://hal.archives-ouvertes.fr/hal-00017169/fr/.

Li, C., Bishas, G., Dale, M. & Dale, P. (2001). *Advances in Intelligent Data Analysis*, Vol. 2189 of *LNCS*, Springer, chapter Building Models of Ecological Dynamics Using HMM Based Temporal Data Clustering – A Preliminary study, pp. 53 – 62.

Mari, J.-F., Haton, J.-P. & Kriouile, A. (1997). Automatic Word Recognition Based on

Second-Order Hidden Markov Models, *IEEE Transactions on Speech and Audio Processing* 5: 22 – 25.

Mari, J.-F. & Le Ber, F. (2003). Temporal and spatial data mining with second-order hidden markov models, *in* M. Nadif, A. Napoli, E. S. Juan & A. Sigayret (eds), *Fourth International Conference on Knowledge Discovery and Discrete Mathematics - Journées de l'informatique Messine - JIM'2003, Metz, France*, IUT de Metz, LITA, INRIA, pp. 247–254.

Mari, J.-F. & Le Ber, F. (2006). Temporal and Spatial Data Mining with Second-Order Hidden Markov Models, *Soft Computing* 10(5): 406 – 414. http://hal.inria.fr/inria-00000197.

Mari, J.-F. & Schott, R. (2001). *Probabilistic and Statistical Methods in Computer Science*, Kluwer Academic Publishers.

Mignolet, C., Schott, C. & Benoît, M. (2007). Spatial dynamics of farming practices in the Seine basin: Methods for agronomic approaches on a regional scale, *Science of The Total Environment* 375(1–3): 13–32. http://www.sciencedirect.com/ science/article/ B6V78-4N3P539-2/2/ 562034987911fb9545be7fda6dd914a8.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B. & Bessières, P. (2002). Mining *Bacillus subtilis* Chromosome Heterogeneities Using Hidden Markov Models, *Nucleic Acids Research* 30(6): 1418 – 1426.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*, Morgan Kaufman.

Pieczynski, W. (2003). Markov models in image processing, *Traitement du signal* 20(3): 255–278.

Rabiner, L. & Juang, B. (1995). *Fundamentals of Speech Recognition*, Prentice Hall.

Schaller, N., Lazrak, E.-G., Martin, P., Mari, J.-F., Aubry, C. & Benoît, M. (2010). Modelling regional land use: articulating the farm and the landscape levels by combining farmers' decision rules and landscape stochastic regularities, Poster session, European Society of Agronomy. Agropolis2010, Montpellier.

Schumm, S. (1977). *The fluvial system*, Wiley, New York. 338p.

Tou, J. T. & Gonzales, R. (1974). *Pattern Recognition Principles*, Addison-Wesley.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley.

LAZRAK, EG (2012). « Fouille de données stochastique pour la compréhension des dynamiques temporelles et spatiales des territoires agricoles. Contribution à une agronomie numérique des territoires ». Thèse de doct. Université de Lorraine.

L'agriculture est l'activité humaine qui utilise et transforme la plus grande partie de la surface terrestre. Son intensification et son uniformisation ont engendré plusieurs problèmes écologiques et environnementaux. Comprendre les dynamiques passées et actuelles des territoires agricoles à des échelles régionales, compatibles avec les échelles où s'expriment les services environnementaux et écologiques, est nécessaire pour mieux gérer l'évolution future des territoires agricoles. Pourtant, la plupart des travaux qui ont étudié les dynamiques agricoles à des échelles régionales ne distinguent pas les dynamiques liées au fonctionnement régulier de l'activité agricole de celles liées à des changements dans son fonctionnement. Les autres travaux rapportés dans la littérature qui font cette distinction présentent toutefois l'inconvénient d'être difficilement reproductibles. Cette thèse vise ainsi à développer une méthode générique de modélisation des dynamiques passées et actuelles de l'organisation territoriale de l'activité agricole (OTAA). Nous avons développé une méthode de modélisation stochastique fondée sur des modèles de Markov cachés qui permet de fouiller un corpus de données spatio-temporelles d'occupations du sol (OCS) en vue de le segmenter et de révéler des dynamiques agricoles cachées. Nous avons testé cette méthode sur des corpus d'OCS issus de sources variées (relevés de terrain, télédétection) et appartenant à deux territoires agricoles de dimensions régionales : le site d'étude de Chizé (430 km^2, Poitou-Charentes) et le bassin versant du Yar (60 km^2, Bretagne). Cette méthode apporte 3 contributions à la modélisation de l'OTAA : (i) la description de l'OTAA suivant une approche temporo-spatiale qui identifie des régularités temporelles, puis les localise en segmentant le territoire agricole en zones compactes de régularités temporelles similaires ; (ii) la fouille des voisinages des successions d'OCS et de leurs dynamiques ; (iii) l'articulation des régularités révélées par notre approche de fouille de données à l'échelle régionale avec des règles identifiées par des experts en agronomie et en écologie à des échelles plus locales en vue d'expliquer les régularités et de valider les hypothèses des experts. Nous avons testé la généricité de la première contribution sur les deux territoires d'études. Les deux dernières contributions ont été développées et testées sur le site d'étude de Chizé. Nos résultats valident l'hypothèse que l'OTAA se prête bien à la représentation par un champs de Markov de successions. Cette thèse ouvre la voie à une nouvelle approche de modélisation de l'OTAA explorant le couplage entre régularités et règles, et exploitant davantage les outils d'intelligence artificielle. Elle constituerait les prémices de ce qui pourrait devenir une agronomie numérique des territoires.

Mots-clés : système de culture, succession de cultures, modèle de Markov caché hiérarchique, agronomie des territoires

---

Agriculture is the human activity that uses and transforms most of the Earth's surface. Agricultural intensification and simplification have created numerous ecological and environmental problems. To better manage the future development of agricultural landscapes, it is important to understand the past and current dynamics of agricultural landscapes at regional scales compatible with the scales where environmental and ecological services manifest themselves. Yet, most studies on agricultural dynamics at regional scales do not distinguish between the dynamics related to a steady-state farming activity and the dynamics related to changes in the mechanisms of farming activity. Meanwhile, studies reported in the literature that make this distinction have the disadvantage of being difficult to duplicate. The purpose of this thesis is to develop a generic method for modelling the past and current dynamics of Landscape Organization of Farming Activity (LOFA). We developed a stochastic modelling method based on Hidden Markov Models that allows data mining within a corpus of spatio-temporal land use data to segment the corpus and reveal hidden agricultural dynamics. We applied this method to land use corpora from various sources (field surveys, remote sensing) belonging to two agricultural landscapes of regional dimension : the study site of Chizé (430 km^2, Poitou-Charentes, France) and the catchment area of Yar (60 km^2, Brittany, France). This method provides three contributions to the modeling of LOFA : (i) LOFA description following a temporo-spatial approach that first identifies temporal regularities and then localizes them by segmenting the agricultural landscape into compact areas having similar temporal regularities ; (ii) data mining of the neighborhood of land use successions and their dynamics ; (iii) combining of the regularities revealed by our data mining approach at the regional level with rules identified by agronomy and ecology experts at more local scales to explain the regularities and validate the experts' hypotheses. We tested the generic nature of the first contribution on the two study sites. The last two LOFA modelling contributions were developed abd tested on the study site of Chizé. Our results validate the hypothesis according to which LOFA fits well a Markov field of land-use successions. This thesis opens the door to a new LOFA modelling approach that investigates the combining of regularities and rules and that further exploits artificial intelligence tools. This work could serve as the beginning of what could become a numerical landscape agronomy.

Key words : cropping system, crop successions, Hierarchical Hidden Makov Model (HHMM), landscape agronomy